云计算环境下云服务用户并发量的区间预测模型

孟 煜 张 斌 郭 军 闫永明
(东北大学计算机科学与工程学院 沈阳 110819)

摘 要 云计算环境下服务用户并发量的预测是云环境自适应资源调整的重要依据,但传统的单值预测所包含的信息量过少,受并发量不确定性影响明显,所以其不足以支持完备的自适应调整策略制定,因而会引发过多无效的调整动作.针对上述问题,该文提出一种云服务用户并发量区间预测模型,通过预测并发量的区间量化其不确定性.该模型利用梯度下降粒子群优化的支持向量机作为主要预测方法.为了更有效地预测不同类型的并发量,提出了一种基于自相关系数以及功率谱分析的AC-PS并发量特征判定规则,并针对不同特征并发量采取不同的区间构造方法.该文通过一个实例分析该区间预测模型对解决自适应资源调整无效问题的有效性,最后利用对比实验验证预测区间的准确性.结果表明,相对于其它方法文中提出的区间预测模型对各类并发量数据的预测精度均达92%以上,其预测效率有76.11%~96.15%的提升,因此提出的并发量区间预测方法能够为避免自适应资源调整无效问题提供可靠支撑.

关键词 云服务;并发量;区间预测;云环境;资源调整
中图法分类号 TP311 **DOI号** 10.11897/SP.J.1016.2017.00378

Prediction Interval Estimation Model of User Concurrent Requests for Cloud Service in Cloud Environment

MENG Yu ZHANG Bin GUO Jun YAN Yong-Ming
(*School of Computer Science and Engineering, Northeastern University, Shenyang* 110819)

Abstract User concurrent requests prediction is a critical basis for dynamic resources self-adaptive adjustment in cloud computing, but with the characteristics of the uncertainty and frequent variation of concurrent requests, traditional single-value prediction is not enough to support an effective self-adaptive adjustment, thereby producing a variety of invalid self-adaptive actions, which brings in the serious degradation of cloud performance and decrease in cloud effectiveness. In this paper, a prediction interval estimation model of user concurrent requests for Cloud service is proposed to quantify the uncertainty of user requests and the probability of variations. The model adopts SVM to obtain the upper and lower bound of the prediction interval, and employs the PSO to train data for the optimal solution. In order to make the model available for different types of user concurrent requests, a rule AC-PS based on autocorrelation coefficient and power spectrum analysis is put forward, which can judge the features of the requests. On account of AC-PS, different concurrent requests consider corresponding construction method of prediction interval. An example of self-adaptive adjustment using proposed prediction interval estimation model is given to analyze the utility in handling the invalidation adjustment. Furthermore, a comparison

收稿日期:2016-05-30;在线出版日期:2016-09-09.本课题得到国家自然科学基金(61572116,61572117,61502089)、国家关键科技研发基金(2015BAH09F02)、省科技项目攻关项目(2015302002)、中央高校东北大学基本科研专项基金(N150408001,N150404009,N140406002)资助.**孟 煜**,男,1990年生,博士研究生,中国计算机学会(CCF)会员,主要研究方向为云计算、服务预测及优化.E-mail:mengyu@stumail.neu.edu.cn.**张 斌**(通信作者),男,1964年生,博士,教授,博士生导师,中国计算机学会(CCF)高级会员,主要研究领域为云计算、服务计算.E-mail:zhangbin@mail.neu.edu.cn.**郭 军**,男,1974年生,博士,副教授,中国计算机学会(CCF)会员,主要研究方向为云计算、服务计算.**闫永明**,男,1981年生,博士研究生,主要研究方向为云计算与服务性能优化.

experiment is implemented to evaluate the accuracy of the proposed model. The experiment results show that compared with other prediction interval estimation models, the model proposed in this work can promise a better prediction performance that the accuracies are all above 92%, and efficiency is promoted sharply by 76.11% to 96.15%. Therefore the proposed approach can avoid the invalidation of self-adaptive adjustments in resources effectively, and promote the utility of cloud environments without losing the quality of cloud services.

Keywords　cloud service; concurrent requests; prediction interval estimation; cloud environment; resources adjustment

1　引　言

云计算模式强调的按需付费模式要求云环境下的服务系统能够为用户提供不同服务等级(Service Level Agreement,SLA)的服务[1],在初始部署云服务时,为满足云服务的 SLA,云服务环境需要针对特定用户并发量制定最佳虚拟资源配置方案.但是由于云服务用户并发量经常实时变化,初始的资源配置无法使云服务一直保持满足 SLA 的运行状态,因此云服务需要运行时在资源代价最小的情况下动态调整资源配置来保证云服务的 SLA[2].自适应的资源调整能够更加有效地应对云环境的实时变化,从而为云服务系统持续的性能保障提供更加有效的支持[3].

在适应用户并发量变化的自适应资源调整过程中,云服务用户并发量的获取方式对资源调整的有效性具有很大的影响[4].云环境下服务用户并发量的获取方式主要分为两种:一种方法是实时获取云服务用户并发量[5-6],根据当前云服务的用户并发访问情况得到并发量的当前值,由于云环境状态的不断变化,这种方式会具有无法避免的滞后性,容易产生大量的不必要调整.另一种获取方式是采用预测的方法获取[7-8],利用并发量的历史数据通过预测得到未来一段时间内的并发量.这种方式能够支持主动的自适应资源调整,减少不必要调整的数量,为调整的有效性提供保障.因此目前并发量的预测方法是云环境自适应资源调整研究中的重点.

在并发量预测领域中,目前的研究主要分为两类.一类是自回归移动平均法[9]、灰色模型[10]以及排队论[11]等传统的序列分析方法,另一类是模式匹配[12]、基于神经网络[13-14]以及支持向量机[15]等机器学习对并发量进行预测.无论是序列分析方法还是机器学习方法都是针对未来一段时间内与时间点相对应并发量的特定值的预测.将这些方法应用到针对适应并发量变化所做的自适应资源调整问题的解决上,可能导致自适应调整过程的无效调整问题,其主要表现为由于预测不准确导致产生不该有的调整以及由于并发量频繁快速变化导致调整方案在调整动作执行后的短时间内失效两个方面.

产生不该有的调整的原因是由于云服务会受到用户行为、业务内容以及服务时间等因素的影响,导致其用户并发量的产生具有很强的不确定性,因此对于并发量预测的误差是无法避免的,其预测结果不可能很准确.无效的资源调整不仅会导致云环境资源的浪费,而且会引发新的违反 SLA 的问题,反而对云服务产生负面影响.如果在并发量的预测过程中,将并发量预测的不确定性也表示出来,进而在自适应调整中将预测的不确定性考虑在内,则能从根本上解决因为预测误差导致的自适应资源调整失误问题.但是传统的单值预测方法无法表示预测的不确定性,而预测的不确定性可以用一个范围以及概率量化出来.另一方面,由于 SLA 中所界定的是并发量的一个范围,通过 SLA 评价云服务性能不需要单个的数值,所以相对于单个预测值而言如果给出一个范围其准确度会提升.因此在处理并发量不确定性的问题上,对于并发量可能产生的范围预测会比单值预测更有效.

在不考虑并发量不确定性的情况下,即使预测结果非常准确,由于并发量在短时间内的频繁快速变化,会出现调整方案在调整动作执行之后立即失效的问题.这样就需要在预测过程中衡量出并发量在一段时间内的变化,然而仅仅依靠单个并发量预测的数值无法表示并发量在短时间内的频繁快速的变化.而如果用上下界限的形式来表示预测结果,则可以衡量出并发量的变化范围,使自适应调整能够从这个范围中提取并发量变化信息,从而避免因为并发量变化过快而引发的调整失效的发生.

解决以上两种自适应资源调整无效问题都要求预测算法不能只给出一个预测值,需要给出一个并发量的区间及其置信度,为此,本文提出一种云服务用户并发量区间预测模型.该模型能够预测未来一段时间内的并发量真实值可能存在的区间及其置信度.利用此模型能够为解决自适应资源调整无效问题提供有效支持.

为了使区间预测模型适用于各种类型的并发量数据,通过分析服务用户并发量历史数据,根据并发量数据的分布趋势,本文提出基于自相关系数以及功率谱分析的AC-PS(Autocorrelation-Power Spectrum)并发量特征判定规则.利用此模型可将并发量分为平稳型、趋势型和周期型共3种特征类型.为了精确地预测3种特征类型的并发量区间,本文提出一种面向多种并发量类型的基于梯度下降粒子群优化[16]的支持向量机的服务用户并发量区间预测方法,针对不同的特征采取不同的区间构造方式,并通过查找支持向量机的参数的最优解来确保并发量预测区间的准确性,进而提高多种类型并发量预测的效率与灵活性.

本文第2节介绍云服务并发量预测方法以及区间预测方法的相关工作;第3节给出云服务用户并发量区间预测模型及并发量类型判定规则;第4节给出云服务用户并发量区间预测的主要过程及详细算法;第5节通过云环境用户并发量区间预测模型的应用实例分析模型的有效性;第6节通过不同区间预测算法的对比分析云环境用户并发量区间预测算法的准确性;第7节对全文进行总结,并简要介绍下一步研究的主要方向.

2 相关工作

2.1 用户并发量预测方法

目前并发量预测的研究主要集中在两类:一类是通过序列分析的方法预测并发量,Xiong等人[17]根据服务的CPU需求、I/O需求利用开放式排队网络分析预测在线负载;Calheiros等人[9]利用自回归模型预测虚拟机负载作为动态分配资源的依据.Chiang等人[18]利用排队论算法预测应用的并发量来为数据中心的资源分配提供支持.另一类是利用机器学习的方法,Huang等人[19]基于线性回归分析对云环境下资源进行了动态调整,可以使云环境下的应用更好地满足SLA需求;Yang等人[12]提出基于模式匹配方法的负载预测模型并用于云资源的弹性计算中.Jiang等人[20]根据应用的历史和实时的性能数据,基于回归分析建立用户行为和负载的模型,通过分析和评估来描述用户行为和负载特性.

从目前的研究可以看出,在分析预测服务性能判定触发条件时,主要是利用排队论、协同过滤以及回归分析预测并发量,无论是序列分析方法还是机器学习方法都是针对未来一段时间内与时间点相对应的并发量特定值的预测.无法针对并发量的特点给出并发量的不确定性以及变化范围.

2.2 区间预测方法

区间预测是一种将预测目标的固有不确定性利用区间的形式表达的预测方法.在区间预测过程中含有两个关键的不确定性量化方式,置信区间(Confidence Interval,CI)和预测区间(Prediction Interval,PI).置信区间表示在给定预测目标的指定设置中目标概率分布的平均值可能落入的范围.预测区间表示在给定预测目标的指定设置中单个观测值可能落入的范围[21].所以在实际的预测过程中预测区间被认为是表示真实值所在区域的更加有效的表示方式.与置信区间相比,由于预测区间包含了未来更多的不确定性信息,较宽的预测区间表示实际的不确定性较高,反之较窄的预测区间表示实际的不确定性较低.因此自适应资源调节需要谨慎地使用较宽预测区间的结果作为参考,而较窄的预测区间结果则可以被信任的利用.由于区间预测能够表达预测目标的不确定性,所以区间预测被广泛应用于金融[22]、气象[23]、化学[21]、医疗[24]以及电力[25]等诸多领域,但是在云环境用户并发量预测的研究中尚未提及.

传统的区间预测方法主要包括线性回归法[26]、分布假设法[27]、朴素贝叶斯法[28]及分类回归树[29].线性回归将历史数据回归成线性函数,以此来构建预测区间.分布假设模型是预先假设了一个预测误差的分布情况,然后通过计算预测目标的误差来确定预测区间.朴素贝叶斯法通过建立先验分布模型,然后利用基于历史数据的最大化分析构造预测区间.分类回归树则是通过使用信息增益来决定回归树的最优划分,然后以每个叶子节点作为学习模型对目标进行预测及区间划分.这些方法都是先进行单值回归预测再利用置信度分位进行区间构建,所以需要各种不确定的数据分布假设,导致预测区间的精确度不理想.而目前一些更复杂的神经网络的方法被用于区间预测中来估计预测目标的不确定性.基于卡尔曼滤波的神经网络也被用于预测区间

的估计中，其中卡尔曼滤波通过将神经网络的参数作为一个状态向量，训练并更新出一个前馈的神经网络，以此来突破不能在区间预测中产生动态的协方差的限制[30]. 基于神经网络的上下界估计法[31]利用一个双输出的神经网络使预测区间的计算避免了对于历史数据分布规律的假设. 但神经网络方法通常具有预测结果受限于局部极小值、参数调整过量以及漫长的计算时间等限制.

基于支持向量机的方法被很多学者用于不确定性估计中. 支持向量机是一种基于统计学理论的监督学习模型，由 Cortes 和 Vapnik[32]于 1995 年首先提出. 支持向量机构建了一种超平面，由此可以有效地解决非线性及样本数量少等问题，被广泛地应用于数据分类以及回归预测模型的构建中[33]. 支持向量机的核心思想是将输入空间 R^n 非线性地映射到一个高维空间 D 上，从而将低维特征空间的非线性回归问题转化为高维特征空间的线性回归问题. 本文利用支持向量机的回归模型来预测并发量的区间. 由于支持向量机在解决回归问题上的突出表现，很多研究将其作为主要的回归模型用于区间预测当中. 在非线性条件异方差预测过程中支持向量机被用于预测电力价格[34]. 为了确定异方差误差的方差函数，一种半参数混合影响的最小二乘支持向量机被用于置信区间与预测区间的预测中[35]. 然而这些方法都将重点放在提高预测区间对真实值的覆盖率上，但忽略了预测区间的宽度，这样对于云服务用户并发量的预测而言，很容易因为预测区间过宽而导致决策无法准确制定的问题. 本文的目标是利用梯度下降粒子群优化的支持向量机构建一种非参数、分布自由的云服务用户并发量预测模型. 支持向量机使用的是全局极小值，突破了神经网络中局部最小值的限制. 而选择合适的参数会使支持向量机的性能得到很大提升. 为了提高预测效率，本文利用梯度下降粒子群算法对支持向量机的参数做出最优化选择.

3　云服务用户并发量区间预测模型

云服务用户并发量区间预测模型使用支持向量机作为主要的机器学习方法，相对于其它区间预测方法，支持向量机在查找全局最小值、结果重复筛选以及基础函数自动选取等方面有着突出优势. 在确定支持向量机参数过程中，引入粒子群算法查找参数的最优解以降低支持向量机的训练误差. 为了更为准确地预测各种特征的云服务用户并发量的区间，预测模型引入本文提出的 AC-PS 并发量特征判定规则对并发量历史数据分类. 针对不同并发特征类型，在预测模型中选取不同的区间构造方法以提高预测区间的精度.

3.1　模型基本结构

云服务用户并发量区间预测模型结构如图 1 所示. 模型主要分为并发量历史数据区间序列构建以及并发量区间生成两个部分. 并发量历史数据区间序列构建是将一个并发量时间序列划分为两个时间序列，这两个序列分别表示历史并发量的上界及历史并发量的下界. 并发量区间生成是利用生成的并发量历史数据区间序列预测出给定时间内的并发量区间.

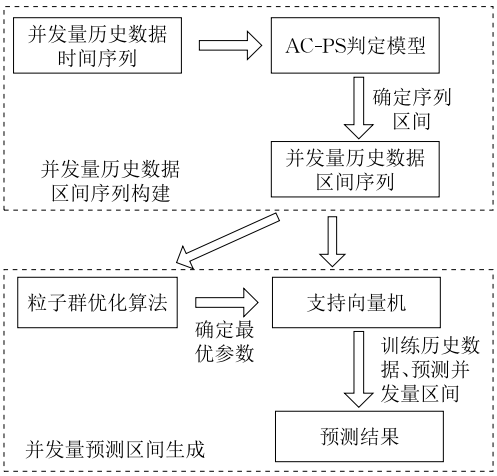


图 1　云服务用户并发量区间预测模型结构

并发量历史数据区间序列构建为其预测区间生成提供数据基础，历史数据区间能够直接影响到预测结果，因此历史数据区间需要准确地包含并发量序列的变化特征. 但是在云环境下不同云服务需要承载的并发量特征是有所不同的，而随着时间的推移，同一服务在不同时间内的并发量也会有显著的差异. 所以需要对并发量历史数据按照其变化特点进行分类，这样才能准确地量化出并发量变化的特征. 为了分析云服务用户并发量特征，根据其数据分布趋势，提出了 AC-PS 判定规则. 此方法将并发量分为平稳型、趋势型和周期型这 3 种类型，进而可以按照并发量的不同类型利用不同的方法构建其历史数据的区间.

对于并发量预测区间的生成，本文选用基于梯度下降粒子群优化的支持向量机作为主要的训练及预测方法. 为了得到两个输出结果构建出预测区间的上下界，需要引入两个支持向量机模型：一个支持

向量机模型利用历史数据区间的上界预测出预测区间的上界;另一个支持向量机模型则用来预测区间下界.两个支持向量机模型都选择径向基函数(Radial Basis Function,RBF)作为其核函数.选择径向基核函数的原因是径向基核函数可以用较少的超参数来表示高维数据,这样就减少了模型选择的复杂度.径向基核函数可以被表示为 $k(x,y)=\exp(-\gamma\|x-y\|^2)$,其中 γ 是核参数.在基于径向基核函数的支持向量机中,支持向量机的惩罚因子 C 与核参数 γ 对于支持向量机的精确度有很大的影响[33].由于预测模型中使用两个支持向量机预测区间的上界和下界,为了使预测区间既能满足预期的置信度又能减小区间宽度,每个支持向量机都需要最优的惩罚因子 C 与核参数 γ,所以模型中需要确定四个参数的数值.

为了获得一个最佳的参数组合使预测区间满足给定的置信度,并在此基础上最小化其预测区间的宽度,需要确定一个合适的参数选择策略.最基础的方法是网格搜索,这种方法是在参数可能存在的范围内顺序的搜索来确定适当的参数.由于并发量的预测需要实时地应用于云环境的资源自适应调整中,所以这种在线预测方式需要对运算的实时性有较高要求.基于网格的最优化查找方法计算复杂度过高,所以非常耗时.这样就需要选择一种优化方法,这里我们选择使用粒子群优化算法作为参数最优解的查找方法.粒子群优化算法是由 Venter 等人[36]首先提出的一种源于对鸟群觅食行为研究分析而产生的群智能算法.为了进一步提高算法的运算效率,在粒子群算法中引入梯度变化思想[16],利用粒子在历史迭代中收敛强度的梯度信息作为其移动速度的更新依据,减少粒子速度及位移的随机性,从而加快粒子群的收敛速度,提高预测算法的实时性.

3.2 基于自相关系数及功率谱的 AC-PS 判定规则

云服务用户并发量是指单位时间内用户对服务的请求数.并发量 con 的定义如下.

定义 1. 并发量.如果用 r 表示服务在 t 时间内的请求数,则并发量 $con=\frac{r}{t}$.

通常情况下,t 以 s 为单位,并发量则为云服务每秒的用户请求数.为了分析云服务用户并发量在时间维度上的特征,需要用到一段时间内相同时间步长下时间点的并发量数据集,也就是并发量时间序列,可如下定义.

定义 2. 并发量时间序列.并发量的时间序列是一个按照时间顺序排列的时间与并发量的二元组的集合,即并发量时间序列

$$S=\{\langle t_1,con_1\rangle,\langle t_2,con_2\rangle,\cdots,\langle t_n,con_n\rangle\}=\{\langle t_i,con_i\rangle\}_{i=1}^{n} \tag{1}$$

其中:n 为并发量个数;con_i 为 t_i 时间的并发量,t_i 需要满足 $t_i<t_{i+1}$.con_i 的大小与定义 1 中 t 的取值相关,t 的取值会影响 con_i 的精度.t 越小 con_i 越接近 t_i 时刻的并发量的真实值,t 越大 con_i 越能代表一段时间内的平均水平.为了减小并发量个别噪声对特征分析造成的影响,本文中 con_i 的取值为时间步长内的平均并发量,即 $con_i=\frac{r_i}{t_i-t_{i-1}}$,其中 r_i 为 t_{i-1} 至 t_i 时间内云服务的用户请求数.

为了在预测过程中判定云服务用户并发量类型,本文提出一种基于自相关系数及功率谱综合判定的 AC-PS 判定规则.AC-PS 判定规则的结构如图 2 所示.

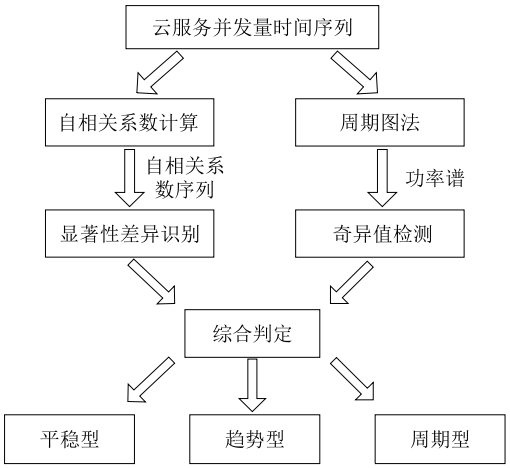


图 2 AC-PS 判定规则结构

自相关系数能够反映时间序列中不同时期观测值之间的相互关系,通过计算不同延迟步长下的自相关系数,可以从不同角度评价云服务用户并发量的特征.由定义 2 可知,$con_1,con_2,\cdots,con_i,\cdots,con_n$ 是过去 n 个时间点对应的云服务用户并发量.将这个序列按照延迟步长为 1 的间隔可化分为 $n-1$ 个二元组:$\langle con_1,con_2\rangle,\langle con_2,con_3\rangle,\cdots,\langle con_i,con_{i+1}\rangle,\cdots,\langle con_{n-1},con_n\rangle$.其延迟步长为 1 的自相关系数为

$$\rho_1=\frac{\gamma(i,i+1)}{\sqrt{Var(con_i)\times Var(con_{i+1})}}=\frac{\sum_{i=1}^{n-1}(con_i-\overline{con_i})(con_{i+1}-\overline{con_{i+1}})}{\sqrt{\sum_{i=1}^{n-1}(con_i-\overline{con_i})^2\sum_{i=1}^{n-1}(con_{i+1}-\overline{con_{i+1}})^2}} \tag{2}$$

其中，$\overline{con}_i=\frac{1}{n-1}\sum_{i=1}^{n-1}con_i$，$\overline{con}_{i+1}=\frac{1}{n-1}\sum_{i=1}^{n-1}con_{i+1}$.

自相关系数的含义不同于两个变量间的相关系数，并发量自相关系数 ρ_1 用来度量并发量时间序列上一时间段并发量观测值与下一时间段并发量观测值之间的关系，从定量的角度来研究上一时间段对下一时间段并发量观测值的影响程度的大小. 自相关系数的延迟步长是组成二元组的两个数值在序列中的间隔数，如果延迟步长为 k，则组成的二元组为 $\langle con_i,con_{i+k}\rangle$，由此可知，并发量时间序列可被划分出 $n-k$ 个二元组，这样可以得到并发量的自相关系数序列.

定义 3.　并发量自相关系数序列. 并发量自相关系数序列是一个由各延迟步长下自相关系数组成的序列 $P_m=\{\rho_1,\rho_2,\cdots,\rho_m\}=\{\rho_k\}_{k=1}^{m}$. 其中 ρ_m 为并发量时间序列在延迟步长为 k 下的自相关系数：

$$\rho_k=\frac{\sum_{i=1}^{n-k}(con_i-\overline{con}_i)(con_{i+k}-\overline{con}_{i+k})}{\sqrt{\sum_{i=1}^{n-k}(con_i-\overline{con}_i)^2\sum_{i=1}^{n-k}(con_{i+k}-\overline{con}_{i+k})^2}}\tag{3}$$

其中，$k=1,2,\cdots,m$. m 为最大延迟步长，一般情况下 $m=\left\lceil\frac{n}{3}\right\rceil$，$\lceil\rceil$ 为向上取整符号.

通过分析并发量时间序列的自相关系数集合 P_m，可以确定并发量时间序列的波动特征. 为了直观地分析并发量序列平稳性特征，选取景点语音导航云服务系统中，景区推荐服务 2016 年 3 月 2 日上午 9:00～9:50 间以 1 min 为步长并发量数据序列. 此序列共包含 51 个并发量值，变化趋势具有平稳型序列的代表性. 其自相关系数序列的延迟步长的取值为 $k=1,2,\cdots,m$，$m=1,2,\cdots,17$，则此用户并发量序列的 P_m 如图 3 所示.

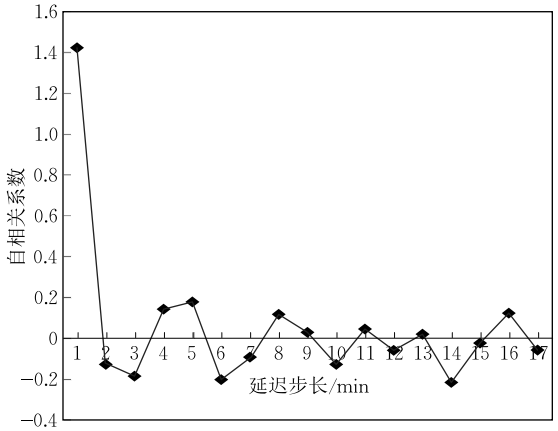


图 3　平稳型云服务用户并发量自相关系数

平稳型云服务用户并发量的第一个自相关系数 ρ_1 为 P_m 中的最大值. 并且 ρ_1 与零有显著性的差异，其余延迟步长下的自相关系数都在零的上下不规则波动，且与零没有显著性差异. 对于趋势型序列，选取有代表性的 GPS 景区识别服务 2016 年 3 月 2 日上午 9:00～11:30 以 2 min 为步长的用户并发量序列. 该序列共有 76 个并发量值，变化趋势为递增，具有明显的趋势型序列代表性. 其自相关系数序列的延迟步长的取值为 $k=1,2,\cdots,m$，$m=1,2,\cdots,25$，用户并发量的 P_m 如图 4 所示.

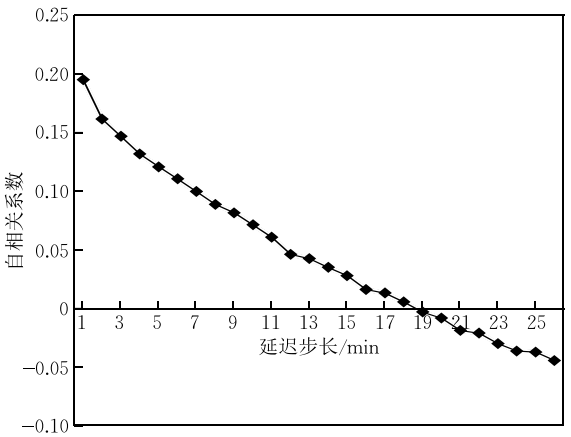
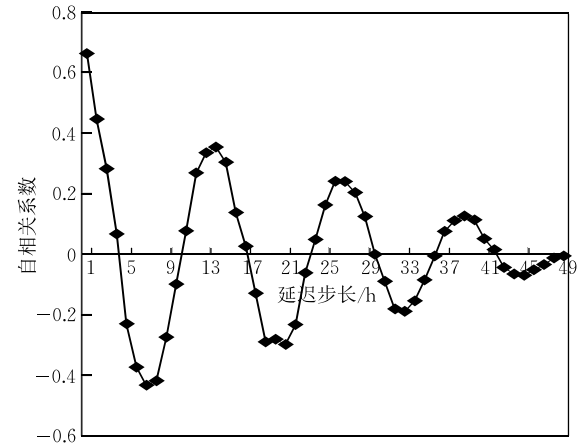


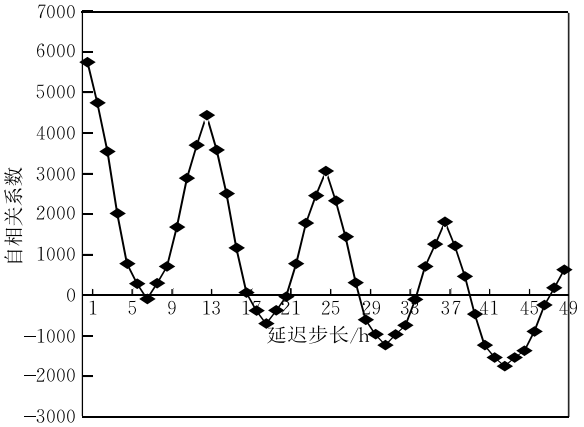
图 4　趋势型云服务用户并发量自相关系数

趋势型云服务用户并发量的第 1 个自相关系数 ρ_1 也为 P_m 中的最大值，但整个 P_m 序列呈单调递减趋势，大多数延迟步长的自相关系数与零有显著性差异. 对于周期型序列，分别选取有代表性的景点图像识别服务与景点路线获取服务 2016 年 3 月 1 日至 3 月 12 日每天早 6:00 到晚 6:00 间步长为 1 h 的并发量序列，每个服务的并发量序列均包含 $13\times12=156$ 个数值，景点图像识别服务的并发量序列在此时间段内有明显的周期性，且趋势平稳，可以代表一般周期型序列. 景点路线获取服务在此时间段内也有明显的周期性，且还具有增长的趋势，可以代表趋势周期型序列，这两种序列的 P_m 如图 5 所示.

周期型云服务用户并发量的自相关系数也具有明显的周期性，规律的交替出现波峰和波谷，并且在波峰至波谷单调递减，波谷至波峰单调递增. 随着延迟步长的增加，每个周期的振幅不断减少. 周期型序列又可分为两种，一般周期型和趋势周期型. 如果并发量数据在某一常数上下周期性波动，则自相关系数会如图 5(a)，在零值上下均匀波动最后趋近于零，这种序列属于一般周期型. 如果并发量周期性的递增或周期性的递减，则自相关系数会如图 5(b)，其自相关系数会呈周期性递减趋势，这种序列属于



(a) 一般周期型并发量自相关系数



(b) 趋势周期型并发量自相关系数

图 5 周期型云服务用户并发量自相关系数

趋势周期型.

由以上 3 种类型并发量时间序列自相关系数分析得知,自相关系数能够明显地区分平稳型序列与趋势型序列.如果时间序列在某一常数上下随机波动,那么其自相关系数序列与零有显著相关性(差异性不显著)[37].

时间序列的功率谱能够通过其谱峰反映时间序列的周期性,本文利用周期图法来估计并发量时间序列的功率谱序列.

定义 4. 并发量功率谱序列. 并发量功率谱序列是一个由各频率下实现序列的功率谱组成的序列 $Q=\{\hat{p}_1,\hat{p}_2,\cdots,\hat{p}_n\}=\{\hat{p}_v\}_{v=1}^{n}$. 其中 $\hat{p}_v$ 为并发量序列频率为 v 的功率谱:

$$\hat{p}_v=\frac{1}{n}|F_v|^2,\ v=1,2,\cdots,n \tag{4}$$

F_v 为离散傅里叶变换,表示为

$$F_v=\sum_{i=1}^{n} con_i W_n^{(v-1)(i-1)},\ v=1,2,\cdots,n \tag{5}$$

W_n 被称为旋转因子,$W_n=e^{-j\frac{2\pi}{n}}$,j 为虚数单位.

由于功率谱序列在 $v=\frac{n}{2}$ 两侧相互对称,所以只分析 $v\in\left[1,\left\lceil\frac{n}{2}\right\rceil\right]$ 上的功率谱即可. 周期型云服务用户并发量功率谱序列 Q 如图 6 所示.

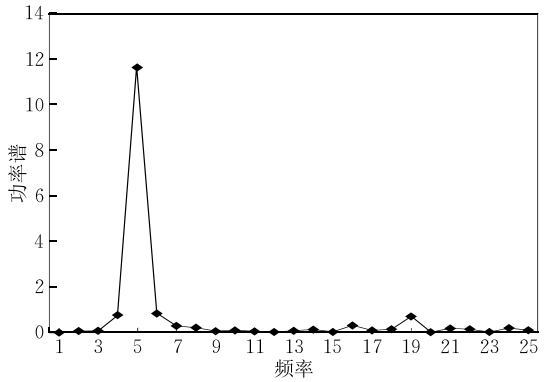


图 6 周期型云服务用户并发量功率谱

从图 6 中不难看出,在周期型云服务用户并发量功率谱序列中包含功率谱突出高的谱峰,这个谱峰则为序列的周期峰.利用功率谱序列可以得到并发量序列的周期.

定义 5. 并发量序列周期数. 并发量序列周期数为周期型并发量序列在一个周期内的数值个数 $n_{\text{period}}=\frac{1}{v_s}n$,其中 v_s 为周期峰对应的频率.

非周期型云服务用户并发量功率谱序列 Q 如图 7 所示.

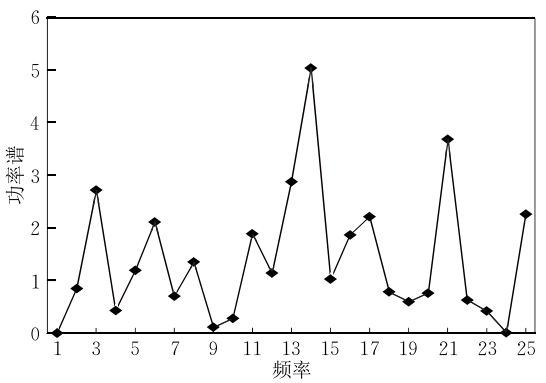


图 7 非周期型云服务用户并发量功率谱

从图 7 中可以看出非周期型云服务用户并发量功率谱序列中没有突出高的谱峰,且各频率下的功率谱大小各异,即无增减性趋势也无规律性.

通过对云服务用户并发量时间序列的自相关系数以及功率谱的分析可知,通过比较自相关系数序列与零值的显著性差异可以判别出并发量时间序列的趋势性,通过检测功率谱序列中是否存在突出峰值可以判别并发量时间序列的周期性. t 检验

(t-Test)[38]是一种最常见检验序列显著性差异的方法，其利用 t 分布可推断出并发量自相关系数序列 P_m 与零发生差异的总体代表性的过错概率 P_ρ. 当 $P_\rho \geq 0.05$ 时，表示自相关系数序列与零差异性不显著；当 $P_\rho < 0.05$ 时，表示自相关系数序列与零差异性显著[38]. 对于功率谱序列突出峰值的识别，本文引入基于 3σ 准则的序列突变奇异值检测方法，将 3σ 准则作用于相邻数据变化序列中，从而找出突出的峰值.

定义 6.　相邻数据变化序列. 相邻数据变化序列为序列相邻数据之差的序列.

则功率谱序列 Q 的相邻数据变化序列

$$H = \{h_2, h_3, \cdots, h_{\lceil \frac{n}{2} \rceil}\} = \{h_v\}_{v=2}^{\lceil \frac{n}{2} \rceil} \tag{6}$$

其中 $h_v = |2\hat{p}_v - (\hat{p}_{v+1} + \hat{p}_{v-1})|$，$v = 2, 3, \cdots, \left\lceil \frac{n}{2} \right\rceil$.

由于当 $v=1$ 时 $\hat{p}_v$ 的值可能会出现突出峰值，当时间序列为趋势型时这个峰值尤为突出，但这个峰值表示整个时间序列只有一个周期，因此这个值不能代表序列的周期性，所以在识别突出峰值时不考虑 $v=1$ 时的功率谱. 如果用 μ_c 代表 H 的平均值，用 σ_c 代表其标准差，则根据 3σ 准则，当 $h_v - \mu_c > 3\sigma_c$ 时，$\hat{p}_k$ 为奇异值，否则为非奇异值. 因此综合以上两种检测方法，可得到云服务用户并发量类型的准确定义.

定义 7.　平稳型并发量序列. 当并发量序列 $P_\rho \geq 0.05$ 且不存在 $\hat{p}_v$ 使 $h_v - \mu_c > 3\sigma_c$ 时称此并发量序列为平稳型并发量序列.

定义 8.　趋势型并发量序列. 当并发量序列 $P_\rho < 0.05$ 且不存在 $\hat{p}_v$ 使 $h_v - \mu_c > 3\sigma_c$ 时称此并发量序列为趋势型并发量序列.

定义 9.　周期型并发量序列. 当并发量序列存在 $\hat{p}_v$ 使 $h_v - \mu_c > 3\sigma_c$ 时称此并发量序列为周期型并发量序列.

4　云服务用户并发量区间预测算法

运用云服务用户并发量区间预测模型对并发量预测的主要过程如下.

过程 1.　云服务用户并发量区间预测主要过程.

(1) 云服务用户并发量分类；

(2) 并发量历史数据区间构建；

(3) 支持向量机参数优化；

(4) 预测并发量区间.

为了给并发量分类提供判定依据，首先需要分别计算并发量历史数据的自相关系数以及功率谱. 自相关系数序列以及功率谱序列包含了并发量序列的特征，需要分别对两个序列进行特征提取. 最后利用提出的 AC-PS 并发量判定规则判定并发量的类型. AC-PS 判定规则需要分别计算云服务并发量时间序列在不同延迟步长下的自相关系数序列以及在不同频率下的功率谱密度序列. 然后分别用两种分析方法对两组序列进行分析. 计算自相关系数序列对于零值的显著性差异，并且对功率谱序列进行奇异值检测. 最后利用两组序列分析结果做综合判定，从而确定并发量的类别. 云服务用户并发量分类具体算法见算法 1.

算法 1.　云服务用户并发量分类算法.

输入：云服务用户并发量历史数据时间序列 S

输出：并发量序列类型 K_{con}，并发量序列周期 period(如果非周期型则输出 0)

```
1.  for i=1; i≤m; i++
2.    {r_i=PS(S,i); //计算延迟步长为 i 的自相关系数
3.      P_m.add(r_i);}
                    //自相关系数加入自相关系数序列中
4.  for i=1; i≤int(length(S)/2); i++
5.    {q_i=PS(S,i); //计算频率为 i 的功率谱
6.      Q.add(q_i);} //功率谱加入自相关功率谱序列中
7.  P_ρ=Ttest(P_m,0); //计算与 0 的差异概率
8.  for i=2; i≤int(length(Q)/2); i++
9.    {h_i=abs(q_i*2-(q_{i-1}+q_{i+1}));
10.     H.add(q_i);} //生成相邻数据变化序列
11. avg=average(H); //H 的平均值
12. s=std(H);
13. period=0; //初始化周期
14. for i=1; i≤length(H); i++
15.   {if h_i-avg>s*3
16.     then{
17.       period=length(H)/i;
18.       K_con='periodType'; //判断为周期型
19.       Break;}}
20. if P_ρ≥0.05
21. then K_con='smoothType'; //判断为平稳型
22. else K_con='tendencyType';//判断为趋势型
23. return K_con, period;
```

并发量历史数据区间构建是将并发量历史序列划分成区间的形式，这样原历史序列 S 被划分成为两个序列，$S^u = \{\langle t_i, con_i^u \rangle\}_{i=1}^n$ 与 $S^l = \{\langle t_i, con_i^l \rangle\}_{i=1}^n$. S^u 是并发量历史数据区间的上界序列，S^l 是并发量历史数据区间的下界序列. 并发量历史数据区间是要将真实值包含在内的，所以并发量的上下界是在原数据序列 S 的基础上依照并发量序列的分类特点采用不同的方式增加及减少得到. 由于这个区间

的生成既要避免预测过程中的不必要的计算，又要防止区间设置过大或过小对于区间特征的干扰，更要使区间能够代表并发量的变化特征，因此区间的构建对于并发量的预测是尤为重要的. 为了体现不同类型并发量的变化特征，每个类型的并发量序列构建区间的方式是不同的. 各种类型并发量的区间构造方式见算法 2.

算法 2. 并发量区间构建算法.

输入：$S, K_{con}, period$

输出：S^u, S^l

1. $S_{\max}=\max(S), S_{\min}=\min(S), S\_len=length(S)$;
2. if $K_{con}==$'$smoothType$' //平稳型并发量序列区间构造方法
3. then{ for $i=1; i\leq S\_len; i++$
4. {$S^u.add(0.5\times(S[i]+S_{\max}))$;
5. $S^l.add(0.5\times(S[i]+S_{\min}))$;}}
6. if $K_{con}==$'$tendencyType$' //趋势型并发量序列区间构造方法
7. then{ for $j=1; j\leq S\_len-1; j++$
8. {$interval.add(abs(S[j+1]-S[j]))$;
9. $interval\_avg=average(interval)$;}
10. for $i=1; i\leq S\_len; i++$
11. { if $tendencytype==$'up'; //递增趋势区间构造方法
12. then{$S^u.add(S[i]+interval\_avg+(S_{\max}-S[i])/abs(S\_len-i))$;
13. $S^l.add(S[i]-interval\_avg)$;}
14. if $tendencytype==$'low' //递减趋势区间构造方法
15. then{$S^u.add(S[i]+interval\_avg)$;
16. $S^l.add(S[i]-interval\_avg-(S[i]-S_{\min})/abs(S\_len-i))$;}}}
17. if $K_{con}==$'$periodType$' //周期型并发量序列区间构造方法
18. then{ for $i=1; i\leq S\_len; i++$
19. { if ($i\leq int(period/2)+1$)
20. then{ for $j=1; j\leq period-1; j++$
21. {$interval.add(abs(S[j+1]-S[j]))$;}
22. $interval\_avg=average(interval)$;}
23. if (($i>int(period/2)+1$) and ($i<S\_len-(int(period)/2-1)$))
24. then { for $j=i-int(period/2)$; $j\leq i+int(period/2)-1; j++$
25. {$interval.add(abs(S[j+1]-S[j]))$;}
26. $interval\_avg=average(interval)$;}
27. if ($i\geq S\_len-(int(period)/2-1)$)
28. then{ for $j=S\_len-period; j\leq S\_len-1; j++$)
29. {$interval.add(abs(S[j+1]-S[j]))$;}
30. $interval\_avg=average(interval)$;}
31. $S^u.add(S[i]+interval\_avg)$;
32. $S^l.add(S[i]-interval\_avg)$;}}
33. return S^u, S^l;

在支持向量机优化参数优化之前需要分别将已经构造出的并发量历史数据区间的上界集合 S^u 与下界集合 S^l 划分成训练集 $S^u_{\text{train}}, S^l_{\text{train}}$ 与测试集 $S^u_{\text{test}}, S^l_{\text{test}}$. 训练集将被用来训练及调整模型，测试集用于支持向量机模型的测试. 支持向量机参数优化主要分为初始化粒子群参数、训练模型、粒子适应度评价、更新局部最优解与全局最优解等过程. 粒子群中每个粒子都包含 4 个元素，也就是两个支持向量机的惩罚因子 C 与核参数 γ. 一般情况下 C 的取值范围为 $[10^{-5}, 10^5]$，γ 的取值范围为 $[0, 10]$. 这两个参数的合适范围不是预知的. 由于在粒子群迭代过程中运用到了梯度下降的方法，所以参数的查找先从两个宽泛的范围开始，根据适应度函数的梯度下降信息决定粒子的移动方向及速度，从而缩小查找范围. 这样做一方面减少了计算量，提高了收敛速度，另一方面保证了全局最优解的覆盖广度以及局部最优解的精确度. 支持向量机优化算法见算法 3.

算法 3. 支持向量机参数优化.

输入：$S^u_{\text{train}}, S^l_{\text{train}}$

输出：$C\_u\_Gbest, Gamma\_u\_Gbest, C\_l\_Gbest, Gamma\_l\_Gbest$ //分别为上下界惩罚因子与核参数的最优解

1. $init(maxParticles, maxIterator, position, velocity)$; //初始化粒子群的最大粒子数、最大迭代次数以及粒子的位置和速度
2. $trainBag\_u, testBag\_u=fold(S^u_{\text{train}}, 5)$; //为上界序列 5 折交叉验证法分包
3. $trainBag\_l, testBag\_l=fold(S^l_{\text{train}}, 5)$; //为下界序列 5 折交叉验证法分包
4. for $i=1; i\leq maxIterator; i++$
5. {$C\_u, Gamma\_u, C\_l, Gamma\_l=PSOtrain(SVM(trainBag\_u, testBag\_u, trainBag\_l, testBag\_l)$; //利用 5 折交叉验证法训练参数
6. $fitness=CWC(C\_u, Gamma\_u, C\_l, Gamma\_l, trainBag\_u, testBag\_u, trainBag\_l, testBag\_l)$; //适应度计算
7. $G=gradient()$; //计算梯度信息
8. if $fitness<fitness\_Pbest$ //如果适应度小于局部最优解
9. then {$fitness\_Pbest=fitness$; //更新适应度局部最优解
10. $C\_u\_Pbest, Gamma\_u\_Pbest, C\_l\_Pbest, Gamma\_$

l_Pbest=*C_u*,*Gamma_u*,*C_l*,*Gamma_l*;}
//更新粒子局部最优解

11. if min(*fitness_Pbest*)<*fitness_Gbest*
12. then {*fitmess_Gbest*=min(*fitness_Pbest*);
//更新适应度全局最优解
13. *C_u_Gbest*,*Gamma_u_Gbest*,*C_l_Gbest*,*Gamma_l_Gbest*=*C_u_Pbest*,*Gamma_u_Pbest*,*C_l_Pbest*,*Gamma_l_Pbest*;}
//更新粒子全局最优解}
14. *position*,*velocity*=*update*(*position*,*velocity*,*G*);
//利用梯度信息更新粒子位置及速度
15. }
16. return *C_u_Gbest*,*Gamma_u_Gbest*,*C_l_Gbest*,*Gamma_l_Gbest*;

在算法 3 中，评价粒子适应度需要一个适应度函数，这个适应度函数既要保证预测区间的精度达到给定的置信度，又要尽可能缩小其区间宽度. 由两个支持向量机组合成的区间预测模型要同时优化预测区间置信度与区间宽度两个优化目标，使用一般的适应度函数是无法准确地评判参数的适应性的. 这里引入一种覆盖宽度评价标准(CWC)[34]作为粒子的适应度评价准则.

定义 10.　覆盖宽度评价标准[34]. 覆盖宽度评价标准为

$$f_{\mathrm{CWC}} = f_{\mathrm{NAW}} \times (1 + \alpha_{\mathrm{CP}} \mathrm{e}^{(-\eta(f_{\mathrm{CP}}-\beta))}) \tag{7}$$

其中，η 和 β 为两个控制 f_{CWC} 规模的超参数，η 是惩罚因子，为一个正数，β 为目标范围覆盖概率，也就是给定的区间置信度. 在训练过程中 $\alpha_{\mathrm{CP}}=1$，在对测试集预测结果的评估中 $\alpha_{\mathrm{CP}}=\begin{cases}0, & f_{\mathrm{CP}}\geqslant\beta\\ 1, & f_{\mathrm{CP}}<\beta\end{cases}$.

CWC 标准是一种综合的评价标准，其涉及到了区间范围覆盖概率 f_{CP} 与标准化区间平均宽度占比 f_{NAW}[34]. 区间范围覆盖概率是真实值落在预测区间内的百分比，能够体现预测区间包含真实值的准确程度. 标准化区间平均宽度占比是区间宽度的平均值与潜在区间宽度的比值，其能够体现区间的宽度.

在粒子群训练过程中，根据定义 7、定义 8 与定义 9，可以得出如下结论.

定理 1.　当 f_{CP} 没有达到目标范围覆盖概率时，f_{CWC} 恒大于 f_{NAW}，且大于的倍数由 f_{CP} 与目标范围覆盖概率 β 的偏差和惩罚因子 η 决定.

证明.　设 $f_{\mathrm{low}}=f_{\mathrm{CP}}-\beta$，当 $f_{\mathrm{CP}}<\beta$ 时，$f_{\mathrm{low}}<0$，$\mathrm{e}^{(-\eta(f_{\mathrm{CP}}-\beta))}>0$，于是有 $\frac{f_{\mathrm{CWC}}}{f_{\mathrm{NAW}}}=1+\mathrm{e}^{(-\eta(f_{\mathrm{CP}}-\beta))}=1+\mathrm{e}^{(-\eta f_{\mathrm{low}})}>1$，并且 $f_{\mathrm{NAW}}>0$，因此 $f_{\mathrm{CWC}}>f_{\mathrm{NAW}}$，且 $\frac{f_{\mathrm{CWC}}}{f_{\mathrm{NAW}}}$ 的数值由 η 与 f_{low} 共同决定.　证毕.

粒子群中的每一个粒子需要被覆盖宽度评价标准重复的评价，由定理 1 可以得出推论 1.

推论 1.　在迭代过程中当 f_{CP} 与 β 差距较大时，粒子的查找首先以提高 f_{CP} 为目的. 随着 f_{CWC} 的减少，f_{CP} 不断接近 β，当 f_{CP} 与 β 差距不大时，f_{CWC} 则会因为 f_{NAW} 的减小而减少，也就是区间的减小. 又因为 f_{CP} 与 f_{NAW} 经常是矛盾的，所以 f_{CWC} 调节 f_{NAW} 与 f_{CP} 平衡的倾向由 η 决定，η 越大最优解的结果越倾向于 f_{CP}，相反 η 越小最优解的结果越倾向于较窄的区间.

由推论 1 在实际操作中 η 一般取值范围为 [10,100]. 这样通过迭代可以保证 f_{NAW} 与 f_{CP} 的平衡，最后随着粒子 f_{CWC} 的最小值的产生得到参数的最优解.

预测区间算法使用支持向量机回归模型分别对并发量区间的上界及下界进行预测. 其中区间上界与区间下界的预测过程完全相同，以区间上界为例，依据定义 2，并发量区间上界时间序列训练样本为 $\{\langle t_i, con_i^U\rangle\}_{i=1}^n$，其支持向量机的回归函数 $f(t_i)$ 可表示为

$$con_i^U = f(t_i) = \langle \boldsymbol{w}, \varphi(t_i)\rangle + b \tag{8}$$

式中：$\boldsymbol{w}$ 是权重向量；$\varphi(x_i)$ 表示从输入空间 R^n 映射到目标空间的非线性函数；$\langle,\rangle$ 是内积符号；b 为偏差. 权重向量 $\boldsymbol{w}$ 与偏差 b 需要从样本训练得出. 利用从梯度下降粒子群优化算法中得到的 *C_u_Gbest*，*Gamma_u_Gbest*，*C_l_Gbest*，*Gamma_l_Gbest* 作用于两个支持向量机，对并发量测试数据的上下界做出预测，最终得出并发量的预测区间. 预测区间算法见算法 4.

算法 4.　预测区间算法.

输入：*C_u_Gbest*,*Gamma_u_Gbest*,*C_l_Gbest*,*Gamma_l_Gbest*,S^u_{train},S^l_{train},S^u_{test},S^l_{test}

输出：Y^u,Y^l

1. *svm_u*=*SVM*(*kernel*='*RBF*',*C*=*C_u_Gbest*,*gamma*=*Gamma_u_Gbest*);
//构造区间上界支持向量机
2. *model_u*=*svm_u.fit*(S^u_{train}); //训练区间上界支持向量机回归模型，得到式(8)的对应关系
3. Y^u=*model_u.predict*($S^u_{\mathrm{test}}.t^u$); //利用区间上界测试集的时间作为输入预测区间上界
4. *svm_l* = *SVM*(*kernel*='*RBF*',*C*=*C_l_Gbest*,*gamma*=*Gamma_l_Gbest*);
//构造区间下界支持向量机

5. $model\_l = svm\_l.fit(S^u_{\text{train}})$; //训练区间下界支持向量机回归模型,得到式(8)的对应关系
6. $Y^l = model\_l.predict(S^l_{\text{test}}.t^l)$; //利用区间下界测试集的时间作为输入预测区间下界
7. return Y^u, Y^l;

5 实例分析

为了表明并发量区间预测模型在解决自适应资源调整无效问题上的效果,引入一个实例分别用并发量单值预测与区间预测来触发自适应调整,通过分析一段时间内的性能与资源消耗的代价/收益,来对比两种预测方法对自适应资源效果的影响,以此来验证区间预测在解决自适应调整无效问题上的有效性.

为了评价云环境下自适应资源调整的收益,需要综合考虑自适应资源调整成本 C_R 与违反 SLA 的补偿代价 P_C.

定义 11. 云环境的综合成本. 云环境的综合成本为云环境资源成本 C_R 与违反 SLA 的补偿代价 P_C 之和.

$$C_E = C_R + P_C \tag{9}$$

由此可知云环境需要通过自适应资源调整来最小化 C_E 从而使收益最大化.

在一个应用实例中,对于一个云服务 A,云环境与其协定的 SLA 可被表示为

(1) 租期 T_R 的租金为 P_R 元.

(2) 在租期内,保证云服务 A 的并发量在区间 $[0, con_m]$ 内的响应时间低于 t_{resp}.

(3) 在租期内,如果响应时间高于 t_{resp},云环境需要为云服务 A 补偿 P_C 元. P_C 的大小由超过 t_{resp} 的时间长度 T_C 而定,具体定义为

$$P_C = \begin{cases} 0, & 0 \le T_C < 10\%\, T_R \\ 5\%\, P_R, & 10\%\, T_R \le T_C < 15\%\, T_R \\ 10\%\, P_R, & 15\%\, T_R \le T_C < 20\%\, T_R \\ 20\%\, P_R, & 20\%\, T_R \le T_C < 25\%\, T_R \\ 35\%\, P_R, & 25\%\, T_R \le T_C < 30\%\, T_R \\ 55\%\, P_R, & 30\%\, T_R \le T_C < 45\%\, T_R \\ 80\%\, P_R, & 45\%\, T_R \le T_C < 50\%\, T_R \\ 100\%\, P_R, & 50\%\, T_R \le T_C \end{cases} \tag{10}$$

本 SLA 的定义参照了 rackspace、Amazon EC2 及阿里云等大型云计算提供商的 SLA 中部分内容. 其中补偿金 P_C 采用根据违约时间比例分段计算的方式定义. 相对于大型云计算提供商 SLA 所针对的长租期而言,本实例中的租期较短. 所以具体的违约时间占比与惩罚比例在大型云计算提供商 SLA 中定义的基础上,根据本实例中的租期总时长略有调整. 云环境为云服务 A 分配资源的最小粒度为 1 个搭载服务 A 的虚拟机副本,所有为云服务 A 分配的虚拟机副本的配置均相同,负载均衡使云服务 A 的用户并发量均分,每个虚拟机副本在保证响应时间少于 t_{resp} 的前提下最多能处理的并发量为 con_v. 云环境自适应调整通过增删云服务 A 的虚拟机副本来平衡服务性能与资源成本. 这样云环境的资源成本就由云环境为云服务 A 开辟虚拟机副本的个数决定,如果一个云服务 A 的虚拟机副本在单位时间 T_{unit} 内的成本为 P_V 元,则云环境的资源成本 C_R 为

$$C_R = \sum_{i=1}^{N} i \frac{T_i}{T_{\text{unit}}} P_V$$
$$\text{s.t.} \sum_{i=1}^{N} \frac{T_i}{T_{\text{unit}}} = T_R \tag{11}$$

其中:i 为云服务 A 使用虚拟机副本的个数,T_i 为云服务 A 占用 i 个虚拟机副本的总时间,N 为租期内云环境为云服务 A 开辟虚拟机副本的最大值.

在自适应资源调整策略中,由于单值预测与区间预测的结果所包含的并发量预测信息量不同,所以触发方式有所不同,因此调整动作也有所差异,最终会导致不同的代价产生. 本文实例中两种预测方法的触发方式均使用某一时间点所对应的预测信息作为触发依据. 规定单值预测的自适应调整的触发方法为:如果并发量预测序列中存在一点 con_j 满足

$$con_j > n_{\text{now}} \times con_v \quad \text{or}$$
$$con_j < (n_{\text{now}} - 1) \times con_v \tag{12}$$

其中 n_{now} 为云环境中正在为云服务 A 提供支持的虚拟机副本数量. 自适应资源调整方案见表 1.

表 1 单值预测自适应资源调整决策方案

条件	调整方案
$con_j > n_{\text{now}} \times con_v$	增加 $\left\lceil \frac{con_j - n_{\text{now}} \times con_v}{con_v} \right\rceil$ 个虚拟机副本
$con_j < (n_{\text{now}} - 1) \times con_v$	减少 $\left\lceil \frac{(n_{\text{now}} - 1) \times con_v - con_j}{con_v} \right\rceil$ 个虚拟机副本

区间预测比单值预测包含了更多的信息,其自适应资源调整触发方法为:如果并发量预测区间上界 con^u_j 与下界 con^l_j 满足

$$\frac{con^u_j - n_{\text{now}} \times con_v}{n_{\text{now}} \times con_v - con^l_j} > 1 \quad \text{or}$$
$$\frac{(n_{\text{now}} - 1) \times con_v - con^l_j}{con^u_j - (n_{\text{now}} - 1) \times con_v} > 1 \tag{13}$$

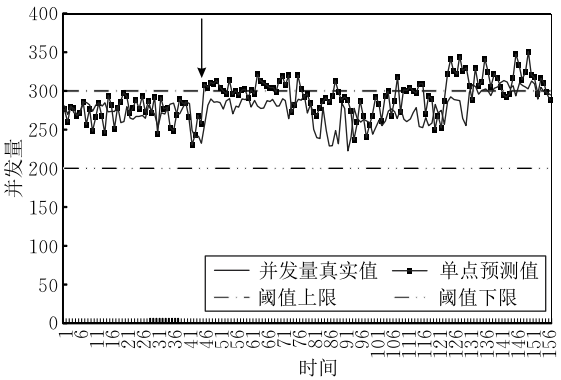
其自适应资源调整方案见表 2.

表 2　区间预测自适应资源调整决策方案

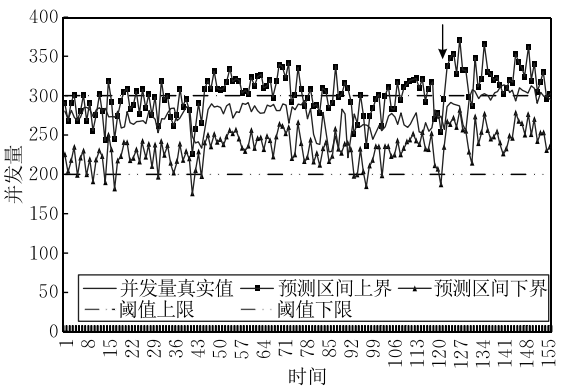
条件	调整方案
$\dfrac{con_j^u-n_{\text{now}}\times con_v}{n_{\text{now}}\times con_v-con_j^l}>1$	增加 $\left\lceil\dfrac{(con_j^u+con_j^l)}{2con_v}-n_{\text{now}}\right\rceil$ 个虚拟机副本
$\dfrac{(n_{\text{now}}-1)\times con_v-con_j^l}{con_j^u-(n_{\text{now}}-1)\times con_v}>1$	减少 $\left\lceil n_{\text{now}}-\dfrac{con_j^u+con_j^l}{2con_v}-1\right\rceil$ 个虚拟机副本

为了进一步分析两种预测方法的效用，在两个真实的云环境资源自适应调整场景中分别对比单值预测与区间预测对云环境总体成本的影响. 两个场景的并发量数据分别来源于景点语音导航云服务系统的景区推荐服务(以下简称服务 A)中两个不同时间段内的用户并发量数据.

场景 1.　云环境中共有 3 个虚拟机副本在为云服务 A 提供支持，即 $n_{\text{now}}=3$. 每个虚拟机副本在保证响应时间少于 t_{resp} 的前提下最多能处理的并发量 $con_v=100$. 利用单值预测与区间预测两种方法分别对同样的 157 个单位时间内的云服务 A 用户统计并发量. 单值预测与区间预测的效果如图 8 所示.



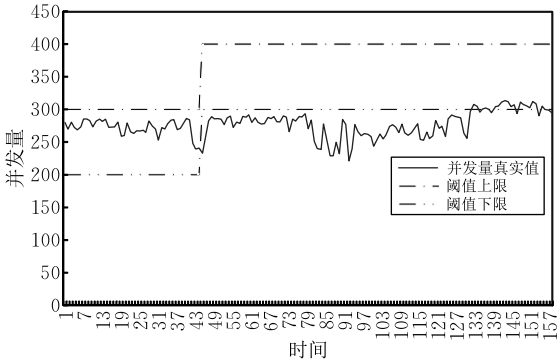
(a) 单值预测值与真实值



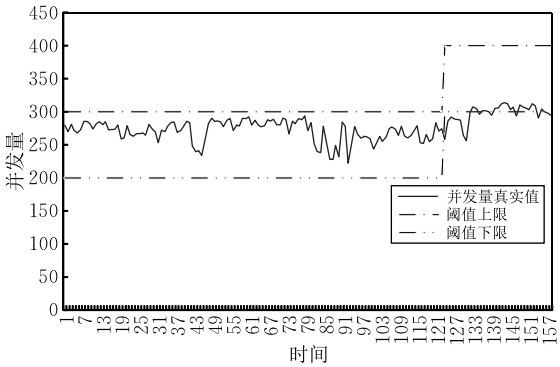
(b) 区间预测值与真实值

图 8　场景 1 中单值预测与区间预测效果

根据式(12)单值预测的触发条件，从图 8(a)中可以看出，从箭头指向的第 46 个时间开始触发自适应资源调整，第 46 个时间对应的数值为 308，即 $con_{46}=308$，代入表 1 中的自适应资源调整方案可得，动态资源调整动作为增加 1 个 A 服务的虚拟机副本. 根据式(13)区间预测的触发条件，从图 8(b)中可以看出，从箭头指向的第 124 个区间开始触发自适应资源调整，第 124 个区间的上界与下界分别是 338 和 268，即 $con_{124}^u=338$，$con_{124}^l=268$，代入表 2 中的调整方案可得，调整动作为增加 1 个 A 服务的虚拟机副本. 通过两种预测方法得出自适应调整效果如图 9 所示.



(a) 单值预测自适应动态资源调整效果



(b) 区间预测自适应动态资源调整效果

图 9　场景 1 中单值预测与区间预测自适应调整效果

由于对比两种算法对自适应调整的效用，两种算法的运行环境一致，所以不考虑自适应动作执行的时间成本及其它干扰因素的影响. 从图 9 中可以看出两种预测方法都有效地避免了违反 SLA 的发生. 计算云环境在这段时间内违反 SLA 补偿代价 $P_C=0$，因此云环境的综合成本为资源成本 C_R，即 $C_E=C_R$. 在这 157 个单位时间内，单值预测的自适应资源调整下云环境在前 42 个单位时间用了 3 个虚拟机副本，从第 43 个单位时间之后用了 4 个虚拟机副本，因此其资源成本 C_R^s 为

$$C_R^s=3\times 42\times P_v+4\times 115\times P_v=586P_v \quad (14)$$

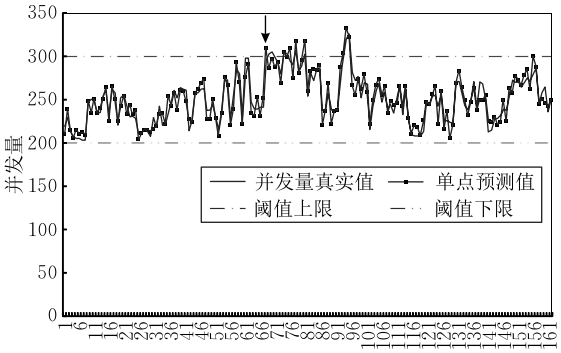
同理，区间预测的自适应资源调整下云环境在前 123 个单位时间用了 3 个虚拟机副本，从第 124 个单位之后开始用了 4 个虚拟机副本，则其资源成本

C_R^d 为

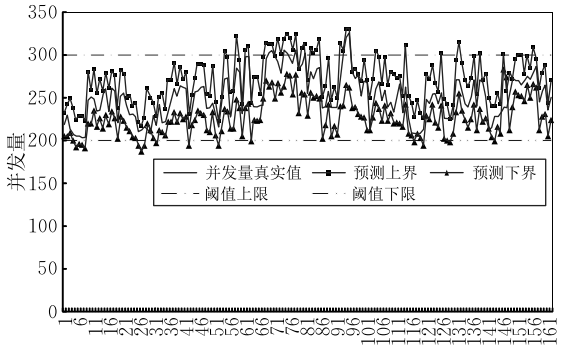
$$C_R^d = 3\times 123\times P_v + 4\times 34\times P_v = 505P_v \quad (15)$$

显然 $C_R^d < C_R^s$，基于区间预测的自适应资源调整的云环境成本明显低于单值预测，区间预测为云环境带来了更高的收益，原因是由于并发量的不确定性导致单值预测不会很准确，从第 43 个单位时间开始预测值高于阈值上限，但是真实值并没有达到阈值上限，这就产生了无效的调整动作，造成成本浪费. 而基于区间预测能够准确地将并发量的不确定性量化成区间形式，使得触发方法可以用真实值出现的概率来判定自适应资源调整的触发，这样能使触发依据更加客观、有效.

场景 2. 云环境中共有 3 个虚拟机副本在为云服务 A 提供支持，即 $n_{\text{now}}=3$. 每个虚拟机副本在保证响应时间少于 t_{resp} 的前提下最多能处理的并发量 $con_v=100$. 利用单值预测与区间预测两种方法分别对同样租期(165 个单位时间)内的云服务 A 用户统计并发量. 单值预测与区间预测的效果如图 10 所示.



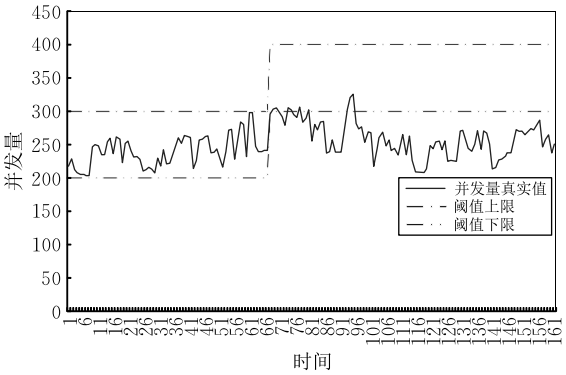
(a) 单值预测值与真实值



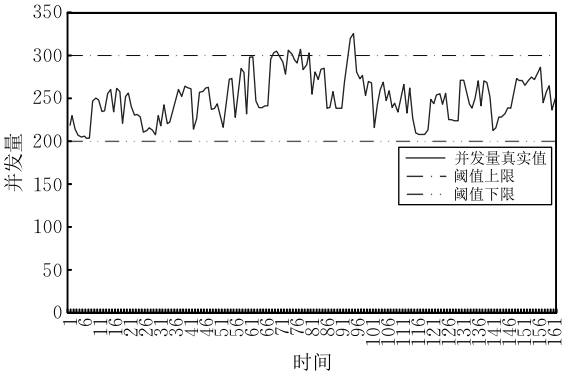
(b) 区间预测值与真实值

图 10 场景 2 中单值预测与区间预测效果

根据式(12)单值预测的触发条件，从图 10(a)中可以看出，从箭头指向的第 69 个时间开始触发自适应资源调整，第 69 个时间对应的数值为 310，即 $con_{69}=310$，代入表 1 中的调整方案可得，调整动作为增加 1 个 A 服务的虚拟机副本. 从图 10(b)中可以看出，区间预测的结果并没有符合触发条件的区间，所以没有任何资源调整动作生成. 通过两种预测方法得出自适应资源调整效果如图 11 所示.



(a) 单值预测自适应动态资源调整效果



(b) 区间预测自适应动态资源调整效果

图 11 场景 2 中单值预测与区间预测自适应资源调整效果

在此场景中也不考虑自适应动作执行的时间成本及其它干扰因素的影响. 从图 11(a)中可以看出，基于单值预测的自适应资源调整在整个时间段内没有违反 SLA 的情况产生. 所以其综合代价为资源代价，$C_E^s = C_R^s$. 单值预测的自适应资源调整策略在 165 个单位时间内，云环境在前 68 个单位时间用了 3 个虚拟机副本，从第 69 个单位时间之后用了 4 个虚拟机副本，因此其资源成本 C_R^s 为

$$C_R^s = 3\times 68\times P_v + 4\times 97\times P_v = 592P_v \quad (16)$$

区间预测没有触发自适应资源调整，因此其资源成本 C_R^d 为

$$C_R^d = 3\times 165\times P_v = 495P_v \quad (17)$$

然而，通过图 11(b)可以看出，基于区间预测的自适应资源调整的云环境在一些时间内云环境违反了 SLA，则需要产生相应的补偿代价，在整个租期 165 个单位时间中云环境总共有 9 个单位时间违反了 SLA，则 $T_R=165$，$T_c=9$，而通过式(10)得出 SLA 的补偿代价 P_C 为 0 元. 这说明短时间的违反

SLA 行为是可以被云服务 A 接受的. 因此区间预测的自适应资源调整的云环境综合成本为 $C_E^d = C_R^d = 495P_v$. 显然，在本场景中基于区间预测自适应资源调整的云环境的综合成本同样低于单值预测. 虽然在区间预测的云环境中有违反 SLA 的情况出现，但只是短暂的，是由于并发量数据的不稳定造成的，这种违反 SLA 的情况云服务 A 是可以接受的. 而在本场景中，单值预测尽管较为准确，但由于其单个数值的限制，无法识别并发量变化，所以导致在一段频繁快速的并发量到来时触发了自适应资源调整策略，增加了虚拟机副本. 但是在频繁快速的并发量持续一段后，并发量又低于之前的阈值上限，这样就造成了调整之后的很长时间内虚拟机副本过剩的情况，导致了成本的增加. 因此相对于单点预测，区间预测能够利用区间优势很好地识别出频繁快速的波动情况，这是因为在相同的置信度内，如果预测值频繁不稳定波动，预测区间会比平稳的预测区间宽，从而表示数据波动的可能性大小.

通过上述两个场景可以得知，区间预测在处理常见的因为并发量不确定性与频繁快速变化而导致的自适应资源调整无效的问题上的效果要优于单值预测.

6　实验及结果分析

6.1　实验数据及环境

本文中所有服务并发量数据均来自于景点语音导航云服务系统(DiTing v3.0)，此云服务系统由东北大学云计算研究实验室开发，其服务端部署于由 20 台物理服务器组成的云环境中，并通过云环境中多个虚拟机副本为该云服务系统提供支持. 景点语音导航云服务系统包含众多子服务，第 3.2 节、第 4 节以及下文对比实验所用到的服务均来自于此系统. 该系统采用 Nginx 作为服务的负载均衡服务器，从中可获取每一个服务的用户并发量数据. 该系统的客户端为手机 App，用户可通过手机客户端对各种服务发出请求.

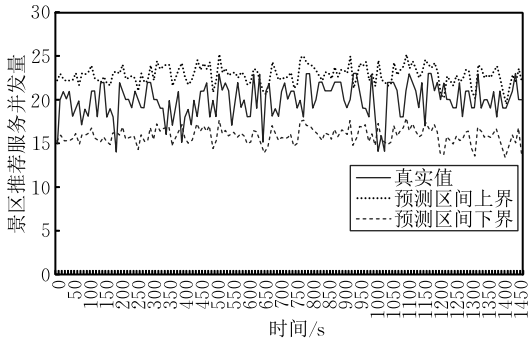
本文中各实验算法均利用 Python 语言编写并在配置为 Intel(R) Core(TM) i7-4790 CPU 3.60 GHz、8 GB 内存、搭载 64 位 Windows 7 操作系统的计算机中运行.

6.2　区间预测准确性对比

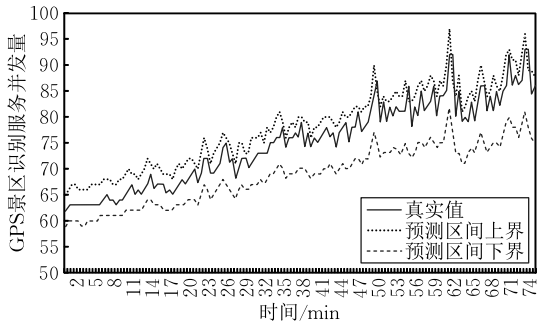
为了进一步验证提出的并发量区间预测模型预测结果的准确性，选择目前主流的 7 种区间预测方法与本文提出的区间预测方法进行对比. 分别利用 3 种不同的区间质量评估方法对比各算法的区间预测效果. 实验数据分别选取景点语音导航云服务系统中景区推荐服务、GPS 景区识别服务以及景点语音讲解服务这 3 个不同特征的云服务并发量. 实验中分别利用线性回归(Linear Legression)[39]、朴素贝叶斯(Naive Bayes)[28]、最小二乘法回归(OLS)[20]、分类回归树(Classification And Regression Tree，CART)[29]、未经过参数优化的支持向量机算法(SVM)[35]、随机森林(Random Forest)[40]、梯度提升决策树(Gradient Boosting Decision Tree，GBDT)[41]以及本文提出的基于用户并发量特征分类的梯度下降粒子群优化支持向量机(ACPS-GPSO-SVM)对各云服务并发量区间做出预测，并比较预测区间的效果. 其中基于线性回归、朴素贝叶斯、最小二乘法、分类回归树以及支持向量机的区间预测模型属于较为传统的区间预测方法，而基于随机森林与梯度提升决策树的区间预测模型法则是目前较为流行且精确度较高的区间预测方法. 分类回归树、支持向量机、随机森林及梯度提升回归树的区间预测精度有赖于其参数的选取，而这些方法的参数主要取决于训练样本的大小及数据变化特征. 为了与本文提出的面向多种特征数据的区间构造方法及基于梯度下降粒子群优化的参数自动选取方法作比较，其它对比算法的参数在预测之前会根据训练数据的规模人为确定为最佳数值.

在实验中，景区推荐服务的并发量选取 2016 年 3 月 5 日上午 9:00～10:40 间以 10 s 为间隔的用户并发量时间序列. 根据定义 1 与定义 2，并发量时间序列中每个时间点所对应的并发量数据为上一时间点到该点的时间段内用户并发量的平均值，即 $con_i = \frac{r_i}{t_i - t_{i-1}}$，其中 r_i 为 t_{i-1} 至 t_i 时间内云服务的用户请求量，$t_i - t_{i-1}$ 为时间序列中相邻点的时间间隔，则在此时间序列中 $t_i - t_{i-1}$ 为 10 s. 此时间序列在 100 min 内总共含有 600 个并发量样本，时间序列中前 450 个样本作为训练样本，后 150 个点作为测试样本，测试样本在训练过程中是不可见的. GPS 景区识别服务选取 2016 年 3 月 5 日至 2016 年 3 月 8 日间每天 9:00～10:15 以 30 s 为时间间隔的平均并发量数据. 同样将时间序列中的 600 个样本分成两组，其中的 150 个数据作为测试样本. 景点语音讲解服务选取 2016 年 3 月 5 日至 2016 年 3 月 14 日间每天 8:00～18:00 时间间隔为 10 min 的平均并发量时间

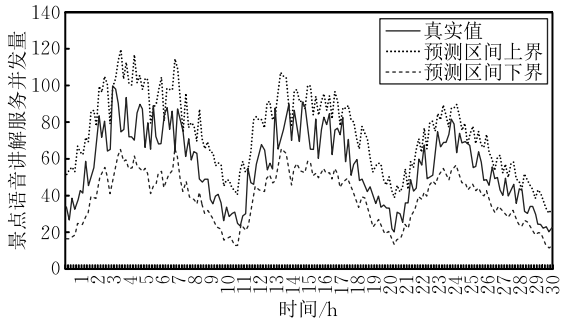
序列.以前7d的数据作为训练样本,后3d的数据作为测试样本.本文提出的ACPS-GPSO-SVM并发量预测模型对以上3种数据的预测结果如图12所示.



(a) 景区推荐服务区间预测效果



(b) GPS景区识别服务区间预测效果



(c) 景点语音讲解服务区间预测效果

图12 ACPS-GPSO-SVM并发量区间预测模型对各服务区间预测效果图

从图12中不难看出,在对3个不同云服务并发量预测过程中,ACPS-GPSO-SVM区间预测模型能够将大部分并发量真实值包含在内,并在此基础上将区间保持较小的范围.有较小数量的真实数据落在了预测区间的外面,但可以从图12(a)与(b)中看出这些超出预测区间的数值存在明显的单独突发性,因为这导致非常短时间的违反SLA,这种单独的峰值在自适应资源调整过程中是可以忽略的.而对于频繁快速波动的情况,如图12(c)所示,并发量在每个波峰达到峰值的时候都会频繁的波动并能持续一段时间,对此ACPS-GPSO-SVM有良好的预测能力,能够将频繁快速的变化包含在区间内,并且用较大的区间代表其变化的可能性,因此从并发量不确定性的度量到频繁快速变化的表达,ACPS-GPSO-SVM都有良好的能力.

为了进一步分析区间预测效果,通过计算各算法结果的区间范围覆盖概率 f_{CP}、标准化区间平均宽度占比 f_{NAW} 与综合覆盖宽度标准 f_{CWC},能够精确地对比出各算法区间预测的精度.区间范围覆盖概率是真实值落在预测区间内的百分比,能够代表预测区间的准确度.标准化区间平均宽度占比是对区间宽度的衡量标准,这个值越大,区间越宽,但是区间的宽度和准确度是矛盾的,大的区间能够有较高准确度,而小区间的准确度会降低,所以在准确度相同的情况下,区间越小说明区间预测的越精确.综合覆盖宽度标准是对前两个评价标准的综合表达,其能够在准确度较低的时候体现出一个较大的数值,而随着准确度的提高,其数值越能趋近于标准化区间平均宽度占比,在达到目标区间范围覆盖概率时 $f_{CWC}=f_{NAW}$.本实验中目标区间范围覆盖概率(给定的目标置信度)β 设置为90%.β 设定为90%的原因是期望的预测区间置信度为90%,这个置信度是一个公认的准确度相对较高的预测区间评价标准,常常作为区间预测的优化目标与评价标准[42].在预测过程中梯度下降粒子群按照 $\beta=90\%$ 的 f_{CWC} 优化支持向量机的参数,同样按照此标准来评价预测的结果.各算法对景区推荐服务并发量区间预测结果的评价见表3.

表3 各算法对景区推荐服务并发量预测区间的评价

算法	f_{CP}/%	f_{NAW}/%	f_{CWC}/%
linear regression	87.33	58.60	280.90
Naive Bayes	90.67	58.82	58.82
OLS	86.67	58.78	364.59
Random Forest	90.00	58.11	58.11
CART	88.00	58.60	217.89
SVM	89.33	58.60	140.38
GBDT	88.67	58.60	172.74
ACPS-GPSO-SVM	92.00	58.50	58.50

景区推荐服务序列具有平稳型的特性,从表3可以看出,每种算法的 f_{CP} 均在86.67%以上,预测区间都具有较高的准确性.其中 f_{CP} 最高的是本文提出的ACPS-GPSO-SVM算法,表明真实值落在其区间的数量最多.对比 f_{CWC} 不难发现,linear regression、OLS、CART、未经参数优化的SVM与GBDT的 f_{CWC} 均明显高于其它3个算法,这是由于这5个算

法结果的 f_{CP} 未达到 β 导致的，而在预测结果的区间范围覆盖概率高于 β 的 3 个算法中，Random Forest 的 f_{NAW} 最小，表明其区间宽度相对较小. 对于平稳型数据的预测，由于其数据的波动范围较小，所以各算法预测区间的宽度较为相近，ACPS-GPSO-SVM 算法的 f_{NAW} 与 Random Forest 只相差 0.39%，而 ACPS-GPSO-SVM 的 f_{CP} 比 Random Forest 高出了 2%，这表明对于平稳型并发量的区间预测，ACPS-GPSO-SVM 算法的预测结果能够在保证区间宽度与其它算法相近的基础上，使其预测区间的准确性达到较高精度. 各算法对 GPS 景区识别服务并发量区间预测的评价结果见表 4.

表 4　各算法对 GPS 景区识别服务并发量预测区间的评价

算法	f_{CP}/%	f_{NAW}/%	f_{CWC}/%
linear regression	87.33	20.73	83.28
Naive Bayes	76.67	16.54	12521.66
OLS	81.33	17.56	8634.91
Random Forest	89.33	20.43	46.20
CART	97.33	17.03	17.03
SVM	91.33	20.80	20.80
GBDT	94.67	19.12	19.12
ACPS-GPSO-SVM	97.33	16.85	16.85

GPS 景区识别服务并发量序列含有递增趋势型特征. 从表 4 中不难看出 f_{CP} 最高的是 CART 与 ACPS-GPSO-SVM 模型，准确率都达到了 97.33%. 对比这两个算法的 f_{NAW} 可以发现，ACPS-GPSO-SVM 的 f_{NAW} 为 16.85%，比 CART 低，表明 ACPS-GPSO-SVM 的区间宽度更小. 因此虽然准确率一致，但是 ACPS-GPSO-SVM 与 CART 相比，在控制区间宽度上还是占有优势的. 各算法对景点语音讲解服务并发量区间预测的评价结果见表 5.

表 5　各算法对景点语音讲解服务并发量预测区间的评价

算法	f_{CP}/%	f_{NAW}/%	f_{CWC}/%
linear regression	45.56	23.63	1.78838E+11
Naive Bayes	50.56	29.57	8282137325.00
OLS	66.67	24.56	1288543.83
Random Forest	86.67	23.99	142.01
CART	78.33	23.67	9467.07
SVM	67.22	23.57	1003833.64
GBDT	88.89	23.67	74.74
ACPS-GPSO-SVM	92.78	23.16	23.16

对景点语音讲解服务并发量的预测，样本选取了较长时间的数值，其特点是具有周期性，1 天为一个周期. 通过表 5 可以看出，对于周期型序列的区间预测 ACPS-GPSO-SVM 模型无论从 f_{CP} 角度还是 f_{CWC} 的角度分析其相对于其它算法都具有较强优势. 其预测效果好的原因是 ACPS-GPSO-SVM 模型中能够识别出并发量序列的周期，并能够根据每个周期内的特征构建区间，这进一步说明特征分类对区间预测是至关重要的. 同时，还需要注意一点的是在这里未经参数优化的 SVM 算法的预测效果较差，说明在不调整参数的情况下其找不到序列的周期特征，这也进一步印证了参数优化的重要性.

通过对以上云服务用户并发量预测区间的实验与评价指数的分析得知，由于本文提出的 ACPS-GPSO-SVM 模型在生成预测区间的过程中使用基于序列特征分类的区间构建方法以及基于梯度下降粒子群参数最优化的支持向量机，使其在预测准确度与区间宽度控制方面均有较强的优势，能够为各种时延下的各种类型云服务并发量提供较强的区间预测支持.

6.3　区间预测效率对比

由于并发量的预测结果服务于实时的资源自适应调整，所以对预测算法的实时性要求较高，需要对其预测效率进行评估. 影响本文提出的区间预测方法效率最主要的因素在于支持向量机参数优化过程. 对 6.2 节中每组服务并发量数据分别利用网格搜索(GridPSO)[25]、普通粒子群算法(PSO)[34]以及本文使用的梯度下降粒子群算法(GPSO)优化支持向量机的参数并完成区间预测. 各算法优化效果及效率对比见表 6.

表 6　算法优化效果及效率对比

预测数据	优化方法	适应度函数(f_{CWC})最小值	达到最小值的迭代次数	预测时间/s
景区推荐服务并发量	GridPSO	58.10	1358	565.21
	PSO	58.33	58	77.72
	GPSO	58.42	16	18.56
GPS 景区识别服务并发量	GridPSO	15.32	1201	539.32
	PSO	16.17	103	132.47
	GPSO	16.36	24	23.34
景点语音讲解服务并发量	GridPSO	26.79	1783	671.55
	PSO	23.62	135	162.95
	GPSO	23.47	22	25.85

表 6 中对比了利用 3 种参数搜索方式的支持向量机对各服务并发量数据区间预测的适应度函数最小值、达到最小值的迭代次数以及预测时间. 其中适应度函数最小值是在迭代优化中以 f_{CWC} 为适应度函数的最小值，预测时间为完成并发量预测的总耗时. 从表中可以看出，前两组实验中网格搜索的最小值较小，但是其迭代次数与运行时间均远高于两种粒子群算法. 与 PSO 相比，GPSO 在前两组实验中的适应度函数最小值表现较差，但差距不明显，且均

能保证支持向量机区间预测的准确性达到给定的置信度.在景点语音讲解服务实验中的 GPSO 的优化结果要优于其它算法,这有赖于梯度下降这种有向的粒子移动方法.3 组实验中 GPSO 达到最小值迭代次数与预测时间均明显低于其它两种算法,GPSO 比 PSO 预测时间少的主要原因是其迭代次数较少,能够保证其快速收敛的原因是粒子梯度下降移动的引入,所以梯度下降方法在对整体预测过程的加速起到了至关重要的作用.综合来看,虽然网格搜索一定程度上有较强的全局搜索能力且每次迭代的用时较短,但由于其遍历式的无针对性查找导致其迭代次数过多而对预测的效率影响较为严重,不适用于实时系统中.对于 GPSO 与 PSO 而言,两种算法的优化结果相似,但基于 GPSO 优化的预测用时比 PSO 快 76.11%以上,3 组实验的预测时间就能保持在 18.56～25.85 s 之间,更适合于云环境的实时资源自适应调整.

对于支持向量机而言,影响其预测效率的主要因素是输入数据的维度[33],输入数据的维度越多,超平面的非线性变换越复杂,而本文中的支持向量机仅使用时间这一个维度作为输入数据,所以数据量在几百甚至几千范围内的变化对支持向量机效率的影响不明显.在服务并发量变化周期较长,数据量极高的情况下,可以采取较长时间间隔的并发量作为实验数据(如本文中景点语音讲解服务并发量的预测),这样不仅能够提高预测效率而且能够避免个别异常数据对整体趋势的干扰,从而突出序列的变化特征.

7 结 论

本文针对云计算环境下由于传统的单值预测提供信息不足而引发的无效资源自适应调整动作过多的问题,提出一种云服务用户并发量区间预测模型.利用支持向量机作为主要学习方法分别预测并发量区间的上界和下界.为了提高支持向量机预测的精度及优化效率,模型引入了梯度下降粒子群优化算法来在线查找支持向量机参数的最优解.为了使模型适用于各种类型的并发量序列,本文提出了一种基于自相关系数以及功率谱分析的 AC-PS 云服务用户并发量特征判定规则.通过 AC-PS 判定规则可有效地根据并发量序列的数据特征对其分类,从而根据不同类型特征采用不同的并发量序列区间构建方法.通过实验证明,相对于其它方法本文提出的区间预测模型对各类并发量数据的预测精度均达 92%以上,其预测效率有 76.11%～96.15%的提升.在自适应资源调整过程中能够有效地避免无效调节动作的产生,一方面有效地保障了云服务的服务性能,另一方面为降低云环境的运营成本、提高云环境的运转效益提供有力支撑.

由于本文对云环境并发量区间预测的研究目前还处于初始阶段,所以还有很多不足需要进一步完善.在未来的研究中,将进一步优化并发量判定规则及区间预测方法,使其能够对特征更复杂的并发量数据做出更有效的区间预测.

致 谢 各位审稿专家及责任编辑对本文的改进提出了宝贵意见,在此一并致谢!

参 考 文 献

[1] Sun Da-Wei, Chang Gui-Ran, Chen Dong, et al. Profiling, quantifying, modeling and evaluating green service level objectives in cloud computing environments. Chinese Journal of Computers, 2013, 36(7): 1509-1525(in Chinese)
(孙大为, 常桂然, 陈东等. 云计算环境中绿色服务级目标的分析、量化、建模及评价. 计算机学报, 2013, 36(7): 1509-1525)

[2] Salah K, Boutaba R. Estimating service response time for elastic cloud applications//Proceedings of the IEEE International Conference on Cloud Networking. Paris, France, 2012: 12-16

[3] Nallur V, Bahsoon R. A decentralized self-adaptation mechanism for service-based applications in the cloud. IEEE Transactions on Software Engineering, 2013, 39(5): 591-612

[4] Kertesz A, Kecskemeti G, Brandic I. An interoperable and self-adaptive approach for SLA-based service virtualization in heterogeneous cloud environments. Future Generation Computer Systems, 2014, 32(2): 54-68

[5] Maurer M, Brandic I, Sakellariou R. Adaptive resource configuration for cloud infrastructure management. Future Generation Computer Systems, 2013, 29(2): 472-487

[6] Rosa L, Rodrigues L, Lopes A, et al. Self-management of adaptable component-based applications. IEEE Transactions on Software Engineering, 2013, 39(3): 403-421

[7] Prevost J J, Nagothu K, Jamshidi M, et al. Optimal calculation overhead for energy efficient cloud workload prediction// Proceedings of the World Automation Congress. Waikoloa, USA, 2014: 741-747

[8] Yang Jing-Qi, Liu Chuan-Chang, Shang Yan-Lei, et al. A cost-aware auto-scaling approach using the workload prediction in service clouds. Information Systems Frontiers, 2014, 16(1): 7-18

[9] Calheiros R, Masoumi E, Ranjan R, et al. Workload prediction using arima model and its impact on cloud applications' QoS. IEEE Transactions on Cloud Computing, 2014, 3(4): 449-458

[10] Jheng J J, Tseng F H, Chao H C, et al. A novel VM workload prediction using Grey Forecasting model in cloud data center//Proceedings of the 2014 International Conference on Information Networking. Phuket, Thailand, 2014: 40-45

[11] Akbari E, Cung F, Patel H, et al. Incorporation of weighted linear prediction technique and M/M/1 Queuing Theory for improving energy efficiency of Cloud computing datacenters//Proceedings of the IEEE Long Island Systems, Applications and Technology Conference. New York, America, 2016: 1-5

[12] Yang Juan, Huang Zhi-Xing, Gao Yue-Xiang, et al. Dynamic learning style prediction method based on a pattern recognition technique. IEEE Transactions on Learning Technologies, 2014, 7(2): 165-177

[13] Li Peng-Hua, Li Ying-Guo, Xiong Qing-Yu, et al. Application of a hybrid quantized Elman neural network in short-term load forecasting. International Journal of Electrical Power & Energy Systems, 2014, 55(2): 749-759

[14] Sahi S K, Dhaka V S. Study on predicting for workload of cloud services using Artificial Neural Network//Proceedings of the International Conference on Computing for Sustainable Global Development. New Delhi, India, 2015: 331-335

[15] Raghunath B R, Annappa B. Virtual machine migration triggering using application workload prediction. Procedia Computer Science, 2015, 54(3): 167-176

[16] Azizipanah-Abarghooee R, Niknam T, Gharibzadeh M, et al. Robust, fast and optimal solution of practical economic dispatch by a new enhanced gradient-based simplified swarm optimisation algorithm. IET Generation, Transmission & Distribution, 2013, 7(6): 620-635

[17] Xiong Kai-Qi, Perros H. Service performance and analysis in cloud computing//Proceedings of the 2009 Congress on Services-I. Los Angeles, USA, 2009: 693-700

[18] Chiang Yi-Ju, Ouyang Yen-Chieh. Profit optimization in SLA-aware Cloud services with a finite capacity queuing model. Mathematical Problems in Engineering, 2014, 2014(1): 1-11

[19] Huang Chenn-Jung, Guan Chih-Tai, Chen Heng-Ming, et al. An adaptive resource management scheme in cloud computing. Engineering Applications of Artificial Intelligence, 2013, 26(1): 382-389

[20] Jiang Cong-Feng, Xu Xiang-Hua, Zhang Ji-Lin, et al. Resource allocation in contending virtualized environments through VM performance modeling and feedback//Proceedings of the 2011 Sixth Annual Chinagrid Conference (ChinaGrid). Dalian, China, 2011: 196-203

[21] Hosen M A, Khosravi A, Creighton D, et al. Prediction interval-based modelling of polymerization reactor: A new modelling strategy for chemical reactors. Journal of the Taiwan Institute of Chemical Engineers, 2014, 45(5): 2246-2257

[22] Seong N Y, Hin P A, Ching S H. Prediction of future asset prices//Proceedings of the American Institute of Physics Conference Series. Langkawi, Malaysia, 2014: 825-830

[23] Zarnani A, Musilek P. Non-parametric interval forecast models from fuzzy clustering of numerical weather predictions//Proceedings of the IFSA World Congress and NAFIPS Meeting. Edmonton, Canada, 2013: 667-672

[24] Nakaya S, Nakamura Y. Adaptive sensing of ECG signals using R-R interval prediction//Proceedings of the Conference Proceedings: Annual International Conference of the IEEE Engineering in Medicine and Biology Society Conference. Osaka, Japan, 2013: 9-12

[25] Shrivastava N A, Khosravi A, Panigrahi B K. Prediction interval estimation for wind farm power generation forecasts using support vector machines//Proceedings of the International Joint Conference on Neural Networks. County Kerry, Ireland, 2015: 1-7

[26] Andersen P E B, Jensen P N-E, Kousgaard P N. Simple linear regression. EAS Publications, 2014, 66(2): 19-39

[27] Sasaki T, Kondo O. Human age estimation from lower-canine pulp volume ratio based on Bayes' theorem with modern Japanese population as prior distribution. Anthropological Science, 2014, 122(1): 23-35

[28] Nandram B, Yin Jia-Ni. A nonparametric Bayesian prediction interval for a finite population mean. Journal of Statistical Computation & Simulation, 2016, 86(16): 3141-3157

[29] Odgers N P, Holmes K W, Griffin T, et al. Derivation of soil-attribute estimations from legacy soil maps. Soil Research, 2015, 53(8): 881-894

[30] Guan Che, Luh P B, Michel L D, et al. Hybrid kalman filters for very short-term load forecasting and prediction interval estimation. IEEE Transactions on Power Systems, 2013, 28(4): 3806-3817

[31] Kasiviswanathan K S, Sudheer K P. Comparison of methods used for quantifying prediction interval in artificial neural network hydrologic models. Modeling Earth Systems & Environment, 2016, 2(1): 1-11

[32] Cortes C, Vapnik V. Support-vector networks. Machine Learning, 1995, 20(3): 273-297

[33] Claesen M, De Smet F, Suykens J A K, et al. EnsembleSVM: A library for ensemble learning using support vector machines. Journal of Machine Learning Research, 2014, 15(1): 141-145

[34] Shrivastava N A, Khosravi A, Panigrahi B K. Prediction interval estimation for electricity price and demand using support vector machines//Proceedings of the International Joint Conference on Neural Networks. Beijing, China, 2014: 3995-4002

[35] Cheng Qiang, Tezcan J, Cheng Jie. Confidence and prediction

intervals for semiparametric mixed-effect least squares support vector machine. Pattern Recognition Letters, 2014, 40(40): 88-95

[36] Venter G, Sobieszczanski-Sobieski J. Particle swarm optimization//Proceedings of the International Conference on Biomedical Engineering & Informatics. Chongqing, China, 2015: 129-132

[37] Dickey D A, Fuller W A. Distribution of the estimators for autoregressive time series with a unit root. Journal of the American Statistical Association, 1979, 74(366a): 427-431

[38] Wang De-Qing, Zhang Hui, Liu Rui, et al. T-test feature selection approach based on term frequency for text categorization. Pattern Recognition Letters, 2014, 45(11): 1-10

[39] Moon H R, Weidner M. Linear regression for panel with unknown number of factors as interactive fixed effects. Econometrica, 2015, 83(4): 1543-1579

[40] Wager S, Hastie T, Efron B. Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. Journal of Machine Learning Research, 2014, 15(1): 1625-1651

[41] Olinsky A, Kennedy K, Kennedy B B. Assessing gradient boosting in the reduction of misclassification error in the prediction of success for actuarial majors. Csbigs Cases in Business Industry Government & Government Statistics, 2014, 5(1): 12-16

[42] Khosravi A, Nahavandi S, Creighton D. Quantifying uncertainties of neural network-based electricity price forecasts. Applied Energy, 2013, 112(4): 120-129

MENG Yu, born in 1990, Ph. D. candidate. His current research interests include cloud computing, service prediction and optimization.

ZHANG Bin, born in 1964, Ph. D., professor, Ph. D. supervisor. His research interests include service computing and cloud computing.

GUO Jun, born in 1974, Ph. D., associate professor. His research interests include service computing and cloud computing.

YAN Yong-Ming, born in 1981, Ph. D. candidate. His research interests include cloud computing and service optimization.

Background

With the development of cloud computing, virtual cloud resources are widely used in real life, and the problem of wasting resources and interfering service performance faced by large servers have been solved preliminarily. However, the traditional method of cloud resource allocation is static allocation, that is to say the initial amount of resources allocation of the cloud resource is "lifelong". The amount of resources will not be changed during the time cloud services are running.

The research in this paper is a kind of prediction optimization problem of user concurrency requests for cloud services in cloud computing environment which avoids the invalidation of self-adaptive adjustments in resources for cloud services effectively and on the other hand promotes the utility of Cloud environment without losing the quality of cloud services. Most recent works in solving this problem consider how to improve the accuracy of the prediction technology. But, actually, the concurrency requests series of cloud services is known to be stochastic in nature with very high uncertainty. In spite of applying various powerful traditional and advanced prediction techniques, it is quite impossible to eliminate forecasting errors. Therefore, accurate forecast of concurrency request is quite a challenging problem for the cloud environment participants. The difficulty in developing reliable prediction module is to estimate the uncertainties prevailing in different kinds of cloud services and incorporating them appropriately in decision making process.

The commonly applied statistical tools for quantifying the uncertainty in the predictions are confidence intervals and prediction intervals. Compared to the confidence intervals, prediction intervals are more relevant to decision makers as they are more informative about the future. Therefore, a kind of prediction interval estimation model of user concurrent requests for cloud service in cloud environment is proposed in this paper. The propose model can quantify the uncertainty related to forecasts by estimating the ranges of the future user concurrency requests which the tradition single value can't hold. It is to a certain extent, the proposed model avoids the e invalidation of self-adaptive resources adjustments in nature work.

This work was supported by the National Natural Science Foundation of China (61572116, 61572117, 61502089), the National Key Technology R&D Program of the Ministry of Science and Technology (2015BAH09F02), the Provincial Scientific and Technological Project (2015302002), and the Special Fund for Fundamental Research of Central Universities of Northeastern University (N150408001, N150404009, N140406002).