文章编号：2095-2163(2021)08-0001-05 中图分类号：TP302.7 文献标志码：A

云平台下资源需求预测方法的研究

冯丹青，吴智博
（哈尔滨工业大学 计算机科学与技术学院，哈尔滨 150001）

摘 要：在云平台下，为了合理配置资源，有效地避免或减少资源不足及配置过量的问题，本文提出一种改进二次指数平滑的组合预测方法。该方法首先考虑负载需求变化的时间性，设置二次指数平滑方法为基础预测模型，同时在考虑预测精度的前提下，提出了使用误差最小原理来确定权重系数 alpha；其次，使用简单预测方法，移动加权平均法（WMA）作为一种响应式的预测，从而进一步减少突发状态下需求不足的产生。实验结果表明本文提出算法在取得较高的预测精度前提下，能够更好地实现对下一阶段负载需求的预测，同时也降低了预测算法的复杂度。

关键词：负载需求；指数平滑；移动加权平均法；组合预测

A hybrid prediction algorithm based on resource demand in the cloud computing

FENG Danqing, WU Zhibo
(School of Computer Science and Technology, Harbin Institute of Technology, Harbin, 150001, China)

【Abstract】 In cloud computing, it is easy to be in over-provisioning or under-provisioning during the resource allocation process. In order to avoid the inefficient state, such as over-provisioning or under-provisioning, we propose a hybrid predictive technique mainly according to the double Exponential smoothing approach. Firstly, the proposed hybrid approach takes the double ES as the basic predictive technique by considering the varying workload along with the time. Meanwhile, on the premise of improving the accuracy, we present the min error method to determine the weight factor, such as the alpha in the ES. Secondly, we use the WMA method as the reactive predictive technique so as to reduce the state in the sudden load. Finally, the experiments prove that the proposed hybrid predictive technique would achieve the better accuracy. And it would reduce the time and space complexity at the same time.

【Key words】 load prediction; ES; WMA; hybrid prediction

0 引 言

云计算是一种计算模式，通过将需求分布到数据中心构成的资源池中，让用户能够根据需求随时自主租用资源[1-2]。根据服务类型的不同，云计算可以分为 IaaS（Infrastructure as a Service）、PaaS（Platform as a Service）和 SaaS（Software as a Service）[3]。SaaS 即应用层，将云平台上开发的应用程序进行封装，提供给用户使用。在 PaaS 上，企业可以通过对环境的部署来进行应用开发；IaaS 作为基础设施层，可以根据用户需要来提供硬件资源。其中，虚拟化作为云平台应用的核心技术之一，可以用来整合资源，如存储资源、应用软件和网络资源等[4-5]。通常，供应商会给用户提供两种配置机制来选择，即长期预留计划和短期需求计划[6]。然而，在资源管理和分配的过程中，容易出现以下两种情况：一是资源供给不足，不能满足用户的需求，导致 SLA 违约；二是资源存在过度供应，导致资源浪费。因此，结合预测技术来分配资源，可以有效地避免资源分配过度或不足的状态。

目前大部分预测都是根据时间序列来进行预测，根据时间可以将负载预测方法分成短期预测和长期预测[7]。长期预测更适合预留机制，而短期预测更适合确定需求机制，实现资源的按需配置。指数平滑法作为一种简单的预测模型，在预测过程中仅需要之前的观察值和相应参数来更新下一阶段的数据[8]。双指数平滑方法，在单指数平滑的方法上进行了改进，有效地实现了对下一时刻资源需求的预测[9]。实际上，变化的负载需求是一种复杂的时间序列，应从统计的角度来对时间序列深入研究和分析。ARIMA 预测是一种常用的时间预测模型，可以通过差分将非平稳序列转换成平稳序列，然后进行预测[10-11]。然而，这些传统的预测模型在提高预测精度的同时，也增加了自身模型的复杂性。实际上，预测模型不仅需要重视预测精度，也需要降低开

作者简介：冯丹青（1980-），女，博士研究生，主要研究方向：云计算；吴智博（1959-），男，博士，教授，博士生导师，主要研究方向：容错计算、移动计算。

收稿日期：2019-01-17

销[12]。因此,也可以对需求负载进行分析,判断其为周期性负载还是非周期性负载,对不同种类型的负载采取不同的预测技术,将有助于提高负载预测的精度。Press 首先将负载分成周期性和非周期性负载,对不同类型的负载分别进行预测,有效地降低了提出算法的开销,从而实现了降低成本的目的[13]。但是,由于其负载变化的多样性,单一的马尔科夫预测模型由于其自身的特点,无法有效对多样性的负载进行实时反馈。CloudScale 是一种预测方法在实现节能的前提下,增加及时填充法,可以有效减少突发状态下资源配置不足的状态[14]。实际上,组合预测作为一种混合预测方法,可以充分考虑单个预测模型的特点,将不同的预测模型混合使用,从而实现对时间序列的有效分析和预测[15]。

虽然,大部分文献针对预测方法研究了很长时间,也取得了很大进展,但在资源分配的过程中,预测技术仍然存在以下问题需要解决[16]:

(1)减少时间复杂度:在资源分配过程中,预测模型的时间和空间复杂度,应该控制在一个合理的范围内;

(2)提高预测精度:在资源分配过程中,有效的定义样本长度,有助于改善预测模型,提高预测精度。

因此,从资源配置的角度来考虑,本文提出了一种简单的组合预测技术,其目的在于减少开销的同时降低 SLA 违约。从时间序列的角度进行切入,同时考虑其曲线的拟合性,因此组合预测算法主要由基础预测模型和响应式的预测方法两部分组成。其实,对于组合预测而言,重点解决的问题在于合理选择预测模型并确定权重[17]。于是,本文选择二次指数平滑方法作为基础预测模型,来预测工作负载曲线。通常,较好的指数平滑方法会通过预测误差来确定最终合适的权重。实际上,如果在每一次负载预测的过程中,均使用误差最小的原理来确定权重。此时,在预测过程中,设置不同的权重,有助于提高预测精度。响应预测模型是针对突发负载的状态提出的,为了降低突发负载状态下资源分配不足的情况,可以使用 WMA 预测模型来调整工作负载,降低误差,提高预测的准确度。本文中使用组合模型的目的是通过分析负载的特点,考虑其时间性和曲线性,合理的选择简单的预测模型进行预测,在提高预测精度的同时也可以降低预测算法的复杂度。

1 云平台下资源管理框架分析

云平台下资源分配的过程中,很容易因为资源配置不足或过度配置而产生 SLA 违约。本文提出一种简单的组合预测模型,其目标在于用简单的预测方法来实现对下一阶段的负载需求进行快速响应及预测。实际上,通过对负载需求进行简单分析,发现负载的变化具有时间性和突发性[18]。此时,可以根据负载的特点来选择一个合适的预测模型。首先,设置 ES 模型作为基础预测模型。由于其时间性的特点,更加适合快速的预测一部分数据;其次,从提高预测精度的前提下考虑问题,WMA 模型由于其简单性,更加适应作为一种响应模型。同时,也可以从架构的角度来考虑云平台下资源的调度过程。实际上,根据云服务的类型可以分为 SaaS、PaaS 和 IaaS。也就是说,用户在 SaaS 层产生需求,PaaS 层上,代理服务商可以将请求配置在 IaaS 层;在 IaaS 层,可以根据需求提供相应的资源。因此,从资源管理和配置的整体角度来讲,可以根据 MAPE 原理来设计并分析产生框架[19],其将主要分成以下 4 个模块,如图 1 所示。

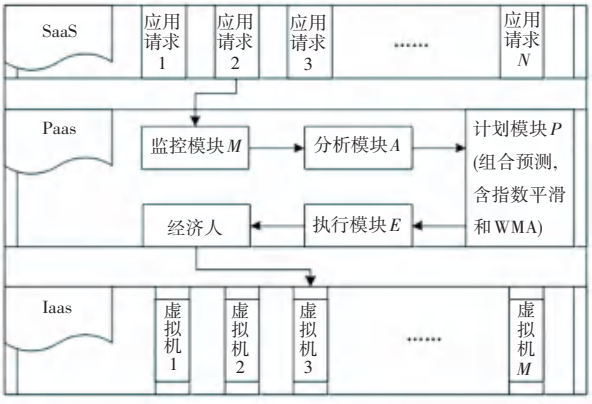

图 1 云平台下资源管理框架

Fig. 1 framework of resource management in the cloud computing

(1)监控模块(Monitor):通过监控模块来查看负载需求,收集原始数据,如负载请求等。利用获取数据从而进一步分析变化的需求曲线。

(2)分析模块(Analyze):监控模块和分析模块均属于数据分析和预处理阶段。该模块分析监控模块收集的信息和数据,为预测行为做准备。预测期间选择指数平滑作为基础预测模型,将选择指数平滑作为基础预测模型。

(3)计划模块(Plan):本文提供的合适的预测方法包括指数平滑和 WMA 预测方法。考虑针对时间序列的预测,同时,为了进一步提高预测精度,使用 WMA 作为一种响应的预测方法,来提高曲线的拟合程度。

(4)执行模块(Execute):在执行模块里,实现组

合预测模型。通过简单的组合模型，可以降低算法复杂度，同时提高预测精度。

2　组合预测模型

在资源分配的过程中，组合预测模型根据负载曲线的特点，对下一阶段曲线进行合理预测。在组合预测过程中，待解决的问题主要有两个：一是合理选择预测模型；二是解决权重的设置问题。为了进一步降低预测算法的复杂性，本文选择简单的时间预测模型，一种改进双指数平滑方法作为基础预测模型。同时，为了减少突发情况的产生，使用 WMA 方法来实现资源负载预测的填充，从而进一步提高预测精度。组合预测模型的分析过程具体如图 2 所示。

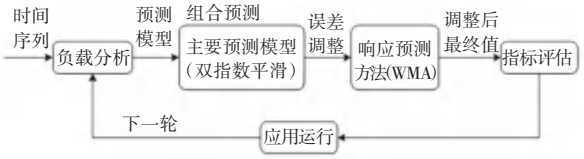

图 2　组合预测

Fig. 2　A hybrid predictive technique

2.1　指数平滑预测

在预测过程中，采用 ES 预测作为基本的预测。同时考虑到进一步的误差修正问题，即 WMA 模型还可以作为响应的预测模型来进一步提高预测精度。也就是说，一次指数平滑预测是指数平滑法，是一种简单的预测曲线，更加适用于解决线性预测的问题。二次指数平滑法在一次指数平滑法的基础上进行改进，能够较好地对曲线进行预测。三次指数平滑法是在二次指数平滑方法上进行预测。虽然三次指数平滑可以改善预测精度，但是由于参数的增加，也增加了该预测方法的复杂度。综上所述，本文选择二次指数平滑法为一种简单基本预测模型，其中需要解决的核心问题为确定 alpha 权重。此时，可以通过误差最小化来确定下一个阶段预测的权重。于是，提出改进的双指数平滑模型具体描述如等式(1)~(4)所示。

$$S_t^{(1)} = \alpha x_{t-1} + (1-\alpha) S_{t-1}^{(1)}, \tag{1}$$

$$S_t^{(2)} = \alpha S_t^{(1)} + (1-\alpha) S_{t-1}^{(2)}, \tag{2}$$

$$y_{t+T} = A_T + B_T T, A_T = 2S_t^{(1)} - S_t^{(2)}, B_T = \frac{\alpha}{1-\alpha}(S_t^{(1)} - S_t^{(2)}), \tag{3}$$

$$\hat{\alpha} = \arg\min_{\alpha} |y_t - x_t|,$$
$$\text{s.t. } y_t = \alpha S_t^{(1)} + (1-\alpha) S_{t-1}^{(2)}, \ \forall \alpha \in [0.1, 0.9]. \tag{4}$$

2.2　组合预测

组合预测模型的主要构成为基础预测模型和响应式预测。考虑到序列的时间性，本文选择二次指数平滑作为预测的主要模型，在每一次预测过程中通过误差最小化原理来确定权重。同时考虑到负载的突发性，使用简单 WMA 预测模型来实现数据的进一步拟合。于是，提出组合预测的具体描述方程(5)。

$$P_{value} = ES_{double}(x_p^{es}) + WMA(|x_t - x_p^w|). \tag{5}$$

组合预测中的核心问题之一是需要解决权重参数，其具体的传统方法有简单加权法、逆平方方差法等。在本文中该预测算法的主旨在于合理选择简单的预测方法，在降低复杂度的前提下，对下一阶段的负载需求进行预测。因此，可以设置组合预测两种方法为等权重。而且本文提出的组合预测算法主要由两部分组成，其具体实现过程描述如下：

(1)首先，考虑到需求负载的时间性，选择二次指数平滑作为基础预测模型，改进的双指数平滑主要通过误差最小原理来确定合理的 alpha 系数；

(2)其次，响应式预测模型采用 WMA 方法。组合预测是为了实现降低复杂度的目标，同时可以提高预测精度。

3　实验结果

为了验证本文算法的有效性，实验采用真实 NASA 负载曲线进行拟合，对比了 3 种预测算法：霍尔特-温特预测、三次指数平滑和双指数平滑。

3.1　实验平台

本文使用 CloudStack 平台管理构建一个数据中心，该数据中心由 7 台物理主机组成。其中，1 台物理主机用来搭建 CloudStack 云平台，另外 6 台物理主机，每台物理主机均安装 Xenserver，在每台物理主机上，均可划分为 3 个 VM，其配置均 1 VCPU，1G memory。每台虚拟机均安装 CentOS 6.9 操作系统。本文采用真实负载 NASA 来验证试验结果，其具体的实现过程主要描述如下：

(1)通过 Jmeter 压力测试工具产生 NASA 数据负载；

(2)监控系统使用 Jmeter 插件来监控一些参数和收集一些信息，如负载需求；

(3)考虑到简单指数平滑方法的特点，可以确定模式匹配长度为最小值，改进后二次指数平滑法可以利用较少的参数来对下一阶段的数据进行预测；

(4)WMA 作为一种简单的预测补偿模型,有效避免资源分配不足的状态。

3.2 指标评估

实际上,预测方法目标之一是在得到较高的预测精度同时误差最小。因此,为了进一步判断预测方法是否符合标准,评价指标可采用平均绝对误差(*MAE*)、均方误差(*MSE*)和均方根误差(*RMSE*)[20]。其中,*MAE* 可以通过计算所有偏差后,再取平均值的方法来获得。,其具体方法表示如等式(6);*MSE* 可以通过计算所有单个实际值和预测值差值平方的平均值方法来获得。其具体方法表示如等式(7);*RMSE* 可以通过计算预测值和真实值的平方和取平均值后求平方根的方法来获得,其具体方法表示如等式(8)。

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |S_i - y_i|, \tag{6}$$

$$MSE = \frac{\sum_{i=1}^{N} (S_i - y_i)^2}{\sum_{i=1}^{N} (S_i)^2}, \tag{7}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (S_i - y_i)^2}. \tag{8}$$

3.3 实验分析

3.3.1 与真实曲线的贴近程度

在组合预测中,合理的选择预测模型是需要面临的挑战之一。考虑到负载需求的时间性和拟合程度,本文选择改进的双指数平滑作为基础预测方法,而简单 WMA 预测方法作为进一步的补偿值预测,从而对突发负载实现较好的曲线拟合。实际上,指数平滑预测方法更适合预测近期的数据值。而 WMA 作为针对拟合性数据的预测,更加适合下一步预测精度的提高。组合预测模型提出的目的之一在于提高预测精度的同时降低算法复杂度。通过对比真实值和实际预测值,可以看出组合预测模型比较贴近真实值。在设置的过程中,通过稍微高一点的预测计划,提高了对突发负载的预测,能够较好的满足预测要求,具体如图 3 所示。

3.3.2 平均绝对误差 *MAE*

误差指标是反映预测精度的一种评估方法,实际上,计算和分析误差的评估指标很多。其中之一为平均绝对误差 *MAE*, 由于其在运算过程中取平均值,使得计算结果能够更加准确。通过对霍尔特-温特,三次指数平滑,二次指数平滑这 3 种预测方法进行对比发现,本文提出的组合预测方法(ESWMA)具有较低的 *MAE* 值,*MAE* 值越小,该预测方法越好,如图 4 所示。

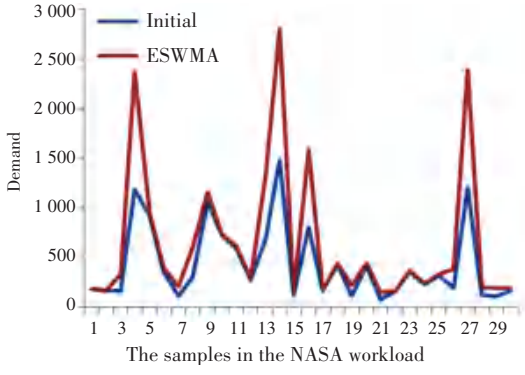

图 3 组合预测和真实情况对比图

Fig. 3 A comparison of the hybrid prediction and actual value

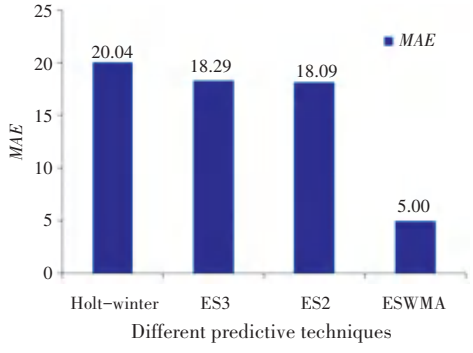

图 4 不同预测方法 *MAE* 值对比

Fig. 4 A comparison of *MAE* evaluation in the different predictions

3.3.3 均方误差 *MSE*

误差指标是反映预测精度的一种评估方法,均方误差 *MSE* 属于其中一种评估指标。均方误差的计算主要用来判断数据的变化程度,计算所得值越小,预测方法越好。通过对霍尔特-温特,三次指数平滑,二次指数平滑这 3 种预测方法对比发现,本文提出的组合预测方法(ESWMA)具有较低的 *MSE* 值, 表明该预测方法较好,也更加符合实际情况,具体如图 5 所示。

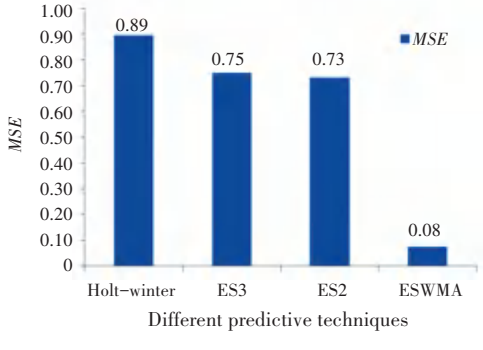

图 5 不同预测方法 *MSE* 值对比

Fig. 5 A comparison of *MSE* evaluation in the different predictions

3.3.4　均方根误差 *RMSE*

误差指标是反映预测精度的一种评估方法，均方根误差 *RMSE* 是一种评估数据变化的指标。其计算作用在于判断数据的变化程度，因此计算所得 *RMSE* 值越小，证明该预测方法越好。通过对霍尔特-温特，三次指数平滑，二次指数平滑这 3 种预测方法来进行对比发现，本文提出的组合预测方法（ESWMA）具有相对很低的 *RMSE* 值，表明该预测方法较好，也更加符合实际情况，如图 6 所示。

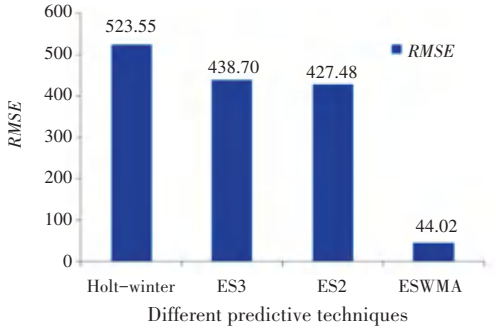

图 6　不同预测方法 *RMSE* 值对比

Fig. 6　A comparison of *RMSE* evaluation in the different predictions

4　结束语

传统的预测模型已经步入发展成熟阶段，可以具有较高的预测精度，但却有着较高的复杂度。实际上，预测技术是一个复杂的计算模型，其面临的挑战之一是提高预测的精度。当然预测精确度越高，参数越复杂，算法复杂度越高。考虑到负载曲线的时间性和突发性，本文设计了一种组合预测模型，利用改进的双指数平滑曲线作为主要预测方法来降低开销，同时考虑到负载的突发性，提出了用 WMA 作为响应式预测模型来进一步提高预测精度，降低误差。在下一步的深入分析和研究中，可以考虑在预测开始之前的工作，包括数据的预处理或是负载预测曲线的进一步分析。

参考文献

[1] RIMAL B P, CHOI E, LUMB I. A taxonomy and survey of cloud computing systems[C]//INC, IMS and IDC, 2009. NCM'09. Fifth International Joint Conference on. Ieee, 2009: 44-51.

[2] ZHOU M, ZHANG R, ZENG D, et al. Services in the cloud computing era: A survey [C]//Universal Communication Symposium (IUCS), 2010 4th International. IEEE, 2010: 40-46.

[3] DINH H T, LEE C, NIYATO D, et al. A survey of mobile cloud computing: architecture, applications, and approaches [J]. Wireless communications and mobile computing, 2013, 13(18): 1587-1611.

[4] WANG L, VON LASZEWSKI G, YOUNGE A, et al. Cloud computing: a perspective study[J]. New Generation Computing, 2010, 28(2): 137-146.

[5] ARMBRUST M, FOX A, GRIFFITH R, et al. A view of cloud computing[J]. Communications of the ACM, 2010, 53(4): 50-58.

[6] CHAISIRI S, LEE B S, NIYATO D. Optimization of resource provisioning cost in cloud computing[J]. IEEE Transactions on Services Computing, 2012, 5(2): 164-177.

[7] 基于广义模糊软集理论的云计算资源需求组合预测研究[J]. 中国管理科学, 2015, 23(5):56-64.

[8] REN X, LIN R, ZOU H. A dynamic load balancing strategy for cloud computing platform based on exponential smoothing forecast [C]//Cloud Computing and Intelligence Systems (CCIS), 2011 IEEE International Conference on. IEEE, 2011: 220-224.

[9] HUANG J, LI C, YU J. Resource prediction based on double exponential smoothing in cloud computing [C]//Consumer Electronics, Communications and Networks (CECNet), 2012 2nd International Conference on. IEEE, 2012: 2056-2060.

[10] BUYYA R, RAMAMOHANARAO K, LECKIE C, et al. Big data analytics - enhanced cloud computing: Challenges, architectural elements, and future directions [C]//Parallel and Distributed Systems (ICPADS), 2015IEEE 21st International Conference on. IEEE, 2015: 75-84.

[11] CALHEIROS R N, MASOUMI E, RANJAN R, et al. Workload prediction using ARIMA model and its impact on cloud applications' QoS[J]. IEEE Transactions on Cloud Computing, 2015, 3(4): 449-458.

[12] AMIRI M, MOHAMMAD-KHANLI L. Survey on prediction models of applications for resources provisioning in cloud[J]. Journal of Network and Computer Applications, 2017, 82(3): 93-113.

[13] GONG Z, GU X, WILKES J. PRESS: PRedictive Elastic ReSource Scaling for cloud systems[J]. CNSM, 2010, 10: 9-16.

[14] SHEN Z, SUBBIAH S, GU X, et al. Cloudscale: elastic resource scaling for multi-tenant cloud systems[C]//Proceedings of the 2nd ACM Symposium on Cloud Computing. ACM, 2011: 5.

[15] 张飞飞. 云资源管理中预测方法的研究与实现[D]. 复旦大学, 2013.

[16] WEINGÄRTNER R, BRÄSCHERG B, WESTPHALL C B. Cloud resource management: A survey on forecasting and profiling models[J]. Journal of Network and Computer Applications, 2015, 47(1): 99-106.

[17] 朱素玲. 组合预测中单项模型选择研究及其权重系数优化[D]. 兰州大学, 2010.

[18] SAPANKEVYCH N I, SANKAR R. Time series prediction using support vector machines: a survey [J]. IEEE Computational Intelligence Magazine, 2009, 4(2).

[19] GHOBAEI-ARANI M, JABBEHDARI S, POURMINA M A. An autonomic resource provisioning approach for service-based cloud applications: A hybrid approach[J]. Future Generation Computer Systems, 2018, 78(1): 191-210.

[20] HWANG R H, LEE C N, CHEN Y R, et al. Cost optimization of elasticity cloud resource subscription policy[J]. IEEE Transactions on Services Computing, 2014, 7(4): 561-574.