云南师范大学学报(自然科学版),2021,41(6):21-24 https://qkgj.ynnu.edu.cn
Journal of Yunnan Normal University(Natural Sciences Edition) DOI:10.7699/j.ynnu.ns-2021-069

基于深度学习的容器云弹性伸缩方法*

徐胜超, 熊茂华
(广州华商学院 数据科学学院,广东 广州 511300)

摘要:通过多指标衡量负载情况,并结合长短时记忆神经网络预测模型,经模型训练和预测后得出负载预测值,共同完成容器云的弹性伸缩决策,避免过度弹性收缩.实验结果证明,该方法可有效全面衡量负载状态;容器云应用负载量的预测值与实际值最大误差值仅为3.2%.

关键词:预测;负载特征;容器云;深度学习;神经网络;弹性伸缩
中图分类号:TP391 文献标志码:A 文章编号:1007-9793(2021)06-0021-04

容器云是一种可以为用户提供资源发布、运行等的应用平台.容器云通过模拟用户资源使用情况,构建虚拟资源处理的相关模型,是云计算中关键技术之一[1-2].互联网应用的频率越来越高,其中产生的数据等信息量十分庞大,容器云的应用效果也会发生变化.为提升容器云使用效果,弹性伸缩必不可少.容器云的弹性伸缩是以Pod为弹性伸缩细胞,1个或几个Pod支持容器云应用工作,实现资源高效利用.

目前,有学者已对容器云资源利用和Kubernetes阈值方法进行研究,如龚坤等人研究了容器云多维资源利用率均衡调度[3]、靳芳等人提出一种网络流量弹性管理方法[4],这些方法中均存在Kubernetes阈值弹性伸缩衡量应用负载局限和伸缩抖动控制不佳的问题.

针对现有方法中存在的弊端,本文采用深度学习方法在原有神经网络基础上完善弹性伸缩,扩展网络训练和网络预测模块,构建新的LSTM(long-term and short-term memory)神经网络模型,实现容器云的有效弹性伸缩.

1 基于LSTM神经网络的容器云弹性伸缩实现

1.1 深度学习的LSTM神经网络

神经网络中数据在不同处理层时,层内未设置节点链接,只是单纯将相邻层进行全链接,此种神经网络不能达到管理负载特征数据的要求[5-6].为了实现负载特征的预测,采用LSTM神经网络完成容器云的负载预测,实现弹性伸缩[7-8].为弥补RNN的缺点,LSTM以RNN为基础,增加一个状态c,扩展结构如图1所示.

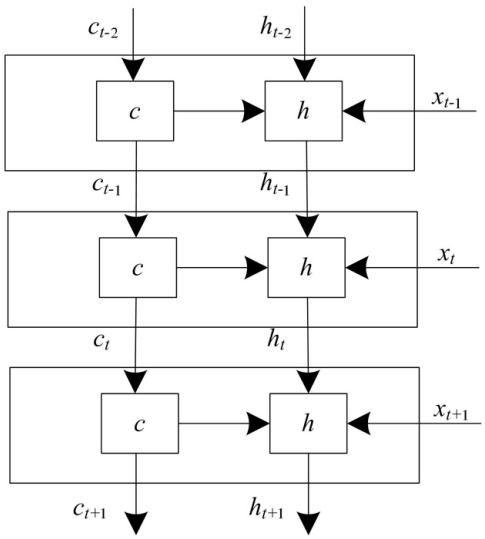


图1 LSTM神经网络扩展结构图

Fig.1 The extended structure of LSTM neural networks

* 收稿日期:2021-07-12
基金项目:广东省高等学校质量工程特色创新基金资助项目(2021RTSCX167);广州华商学院校内导师制科研基金资助项目(2021HSDS15).
作者简介:徐胜超(1980-),男,湖北武汉人,硕士,讲师,主要从事并行分布式处理软件方面研究.
通信作者:徐胜超.E-mail:isdooropen@126.com.

从图 1 可知,该结构中设置了三个对应处理层,以方便计算状态 c、状态 h 和模型输出.

1.2 容器云弹性伸缩实现

根据容器云的负载特征数据和循环神经网络避免复杂的研究准则,预测模型中设置 5 个模块,包括输入层、输出层、中间隐藏层,外加网络训练和网络预测模块,如图 2 所示.

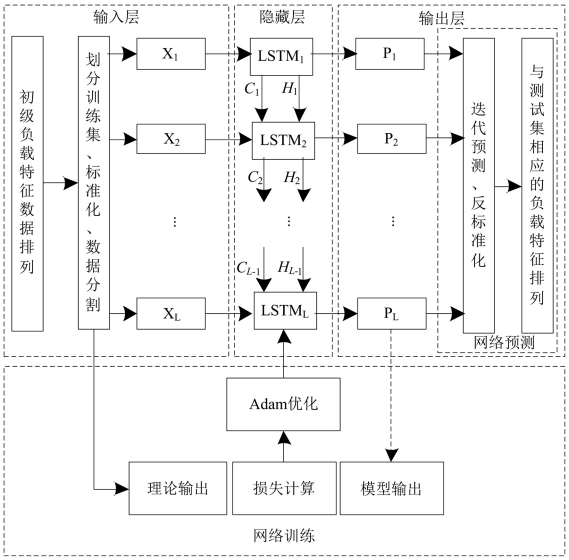


图 2 LSTM 神经网络预测模型

Fig.2 The prediction model of LSTM neural networks

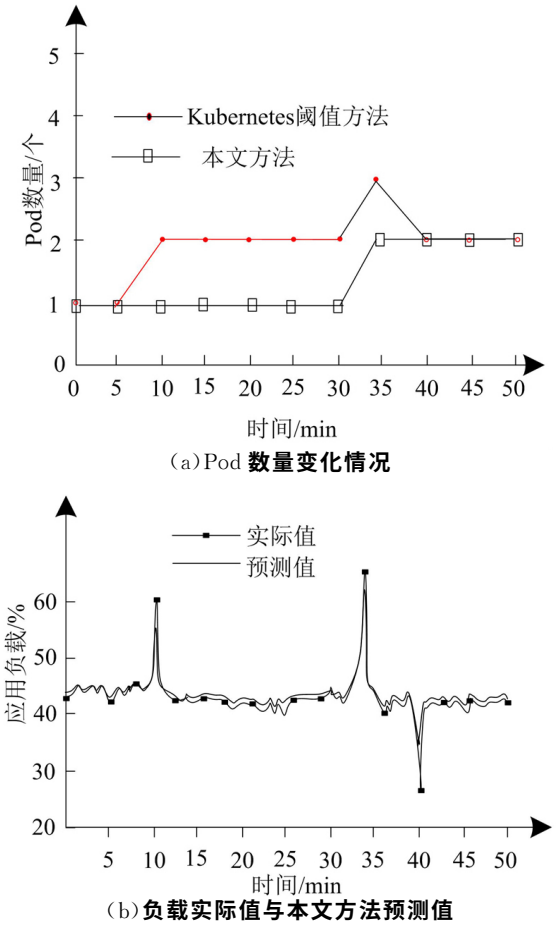
从图 2 中可知,输入层主要管理初级容器云负载特征数据,使其符合网络输入标准.隐藏层通过上扩展的 LSTM 循环链接,建立 LSTM 神经网络.输出层负责呈现预测结果[9].通过 Adam 优化方法训练网络[10-11],从而实现容器云的弹性收缩.

2 实验与性能分析

由于复合型容器化应用负载情况复杂,更能展示本文方法特点,因此实验以 CPU/MEM 复合型容器化应用为例,通过本文方法和 Kubernetes 阈值方法分别对复合型容器化应用实施弹性伸缩.

在复合型容器中设置应用集群,集群中确定一个主控节点和两个一般节点.实验参考阿里 2017[9] 混布云数据中容器化应用资源利用数据将伸缩阈值设成 45%,此时扩容阈值和缩容阈值算得分别为 58%和 32%.

实验初始容器化应用 Pod 数据实例为 1,Kubernetes 阈值方法和本文方法分别进行弹性伸缩的 Pod 数量变化情况见图 3(a)、50 min 内应用负载实际值与本文方法预测值的改变情况见图 3(b).



(a)Pod 数量变化情况

(b)负载实际值与本文方法预测值

图 3 Pod 数量变化及应用负载预测情况

Fig.3 The Pod Variance and prediction results of application loads

图 3 显示,在前 10 min 内容器化应用负载的实际值和预测值均未超过伸缩阈值,此时 Pod 实例数仍旧为 1.10 min 后应用负载出现骤升骤降现象,此时容器化应用负载实际值超过扩容阈值,满足 Kubernetes 弹性伸缩条件,因此实施扩容使 Pod 个数为 2;但预测值低于扩容阈值,未满足本文方法扩容条件,即实际值和预测值没有同时超过扩容值,Pod 个数保持为 1,30 min 内都没有达到需扩容的条件,可见本文方法准确衡量容器化应用负载,避免了扩容.时间到达 35 min 时,容器化应用负载实际值和预测值均大于扩容阈值,本

文方法和 Kubernetes 阈值方法都进行了扩容，Pod 实例数分别为 2 和 3.时间到达 40 min 时，容器化应用负载降值到 40% 以下，低于缩容阈值，Kubernetes 阈值方法立刻对容器化应用进行缩容，Pod 实例数降为 2；由于本文方法启动缩容的另一个条件，即预测值低于缩容阈值没有满足，因此本文方法的 Pod 实例数仍为 2，本文方法对容器化应用少进行一次无必要缩容.综合分析，本文方法同时根据容器化应用负载预测和实际状况对容器化应用 Pod 的弹性伸缩进行控制，能够应对容器云中短时间内发生的应用负载突变，避免无用的伸缩抖动，衡量精准、伸缩谨慎、系统应用性强.

根据上述实验，测试本文方法对容器云应用负载特征预测误差，得到的结果如表 1 所示.

表 1　本文方法的容器云应用负载特征预测误差

Table 1　The container cloud application loads prediction errors of our approach

时间/min	误差/%
5	1.9
10	2.3
15	0.8
20	2.6
25	3.1
30	1.8
35	2.2
40	1.6
45	2.4
50	3.2

根据表 1 可知，50 min 内容器云应用负载量预测值与实际值波动不明显，最大误差值仅为 3.2%.说明本文方法预测容器化应用负载量准确，预测值可用于制定弹性伸缩标准.这是由于本文方法将深度学习运用到容器云弹性伸缩方法中，提高了伸缩方法的准确性.

将本文方法运用到实际容器云平台的资源调度器中，通过资源调度情况反映本文方法的性能.在容器云中，弹性伸缩服务调整稳定所需时间大约 250 s，调度器要随时采集伸缩需求.图 4 为采集容器云中云文档案管理应用负载量，图 5 为本文方法实际运用效果数据.

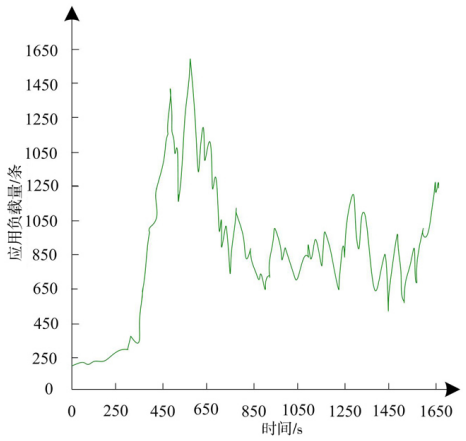


图 4　容器云中云文档案管理应用负载量

Fig.4　The application loads numbers of documents in container based cloud

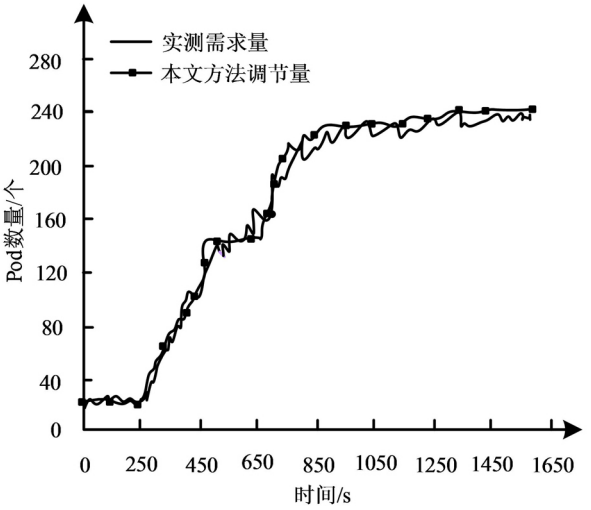


图 5　本文方法实际运行结果

Fig.5　The real testing experimental results of our approach

由图 4、图 5 可知，容器化应用负载量变化巨大，需要及时根据负载变化调节，保证系统顺利运行，而本文方法能够根据负载特征预测进行弹性伸缩，图 5 中可看出本文方法符合实际需求，并且更加稳定，可以避免多余弹性伸缩.

3　结语

针对现有方法中容器云弹性伸缩方法中存在的不足，本文研究的基于深度学习的容器云弹性伸缩方法，通过全面衡量负载特性并对应用类型进行分析，利用深度学习方法中 LSTM 神经网络

预测模型完善预测功能,使应用负载预测更准确,同时可根据负载预测值和实际值双标准控制伸缩动作.经实验证实,本文研究方法具有预测精准、伸缩高效和应用性强的优势.

参考文献:

[1] A hybrid genetic programming hyper-heuristic approach for online two-level resource allocation in container-based clouds[C].2019 IEEE Congress on Evolutionary Computation(CEC),Wellington,New Zealand,June 10-13,2019.

[2] A genetic programming hyper-heuristic approach for online resource allocation in container-based clouds[C].AI 2018:Advances in Artificial Intelligence,31st Australasian Joint Conference,Wellington,New Zealand,December 11-14,2018.

[3] 龚坤,武永卫,陈康.容器云多维资源利用率均衡调度研究[J].计算机应用研究,2020,37(4):148-152.

[4] 靳芳,龙娟.一种面向 Kubernetes 集群的网络流量弹性管理方法[J].北京交通大学学报,2020,44(5):81-90.

[5] 欧阳红兵,黄亢,闫洪举.基于 LSTM 神经网络的金融时间序列预测[J].中国管理科学,2020,28(4):30-38.

[6] 杨青,王晨蔚.基于深度学习 LSTM 神经网络的全球股票指数预测研究[J].统计研究,2019,36(3):67-79.

[7] 邵必林,王莎莎.基于负载预测的 HDFS 动态负载均衡改进算法[J].探测与控制学报,2019,41(2):77-82.

[8] 梁毅,曾绍康,梁岩德,等.一种基于周期性特征的数据中心在线负载资源预测方法[J].计算机工程与科学,2020,42(3):381-390.

[9] 林涛,冯竞凯,郝章肖,等.基于组合预测模型的云计算资源负载预测研究[J].计算机工程与科学,2020,42(7):1168-1173.

[10] 曾旭禹,杨燕,王淑营,等.一种基于深度学习的混合推荐算法[J].计算机科学,2019,46(1):133-137.

[11] 谷远利,李萌,芮小平,等.基于深度学习的网约车供需缺口短时预测研究[J].交通运输系统工程与信息,2019,19(2):227-234.

The Elastic Scaling Method of Container Cloud Based on Deep Learning

XU Sheng-chao,XIONG Mao-hua

(School of Date Science,Guangzhou Hua Shang College,Guangdong,Guangzhou,511300,China)

Abstract:According to the load characteristics of different application types,the load is measured by multiple indicators.Combined with the long-term and short-term memory neural network prediction model,the load prediction value is obtained after model training and prediction,and the elastic scaling decision of container cloud is jointly completed to avoid excessive elastic shrinkage.The experimental results show that our method can measure the load state effectively and comprehensively.The maximum error between the predicted value and the actual value of container cloud application load is only 3.2%.

Keywords:Forecast;Load characteristics;Container cloud;Deep learning;Neural network;Elastic expansion