Signal, Image and Video Processing (2023) 17:2101–2109
https://doi.org/10.1007/s11760-022-02424-8

ORIGINAL PAPER

A cloud computing load prediction hybrid model with adaptive weight strategy

Chao Xiong[1] · Yepeng Guan[1,2]

Received: 28 March 2022 / Revised: 31 October 2022 / Accepted: 28 November 2022 / Published online: 29 December 2022

Abstract
In recent years, cloud computing has revolutionized computing. The resource usage prediction of machine workloads has been one of the most important and challenging problems. Load prediction in cloud computing is challenging due to the inefficiency of feature extraction and complex loading environment. Attention mechanism based Multi-level Feature Extraction (MFE) model has been developed. Empirical mode decomposition has been selected to extract the sequence surface features. Multi-level LSTM has been used to extract multi-level features of the decomposed load in the empirical mode decomposition. Attention mechanism has been adopted to get different weights for the extracted features. In order to further get causal relationship of time series data, Empirical mode decomposition based Temporal Convolutional Network (ETCN) model has been developed. Multiple temporal convolutional network residual blocks are stacked to improve the historical memory ability and increase the local receptive field. An adaptive weight strategy has been proposed to assign different weights to the developed MFE and ETCN. The weights are updated adaptively according to the errors in MFE and ETCN, respectively. The proposed method has excellent performance in cloud computing on some challenging datasets by comparison with some state-of-the-art ones. The proposed method achieves an average improvement of 30.23% and 41.5% on the NASA datasets and Saskatchewan datasets, respectively.

Keywords Cloud computing load prediction · Multi-level feature extraction model · Empirical mode decomposition based temporal convolutional network model · Adaptive weight update

1 Introduction

Cloud computing is characterized with high robustness, higher scalability and on-demand services, more and more organizations are inclined to deploy their applications on cloud. Among parallel and distributed data, researchers have done a lot of work. A dual-phase pipeline task scheduler (D2PTS) [1] and a Parallel Random Forest (PRF) [2] algorithm are proposed. The utilization rate of resources are effectively improved. The data distribution of distributed parallel environment and task scheduling mechanism are improved. The CPU-GPU heterogeneous clusters provide more computing power than the CPU cluster, and effectively integrate the GPU into the distributed processing framework, the computing power of the cluster would be greatly improved [3]. For large-scale time series data, reducing data scale, extracting core information and improving algorithm performance are also the research directions [4]. Load forecasting plays an important role in cloud computing, edge computing and resource allocation [5].

Users can get high quality, strong security and highly scalable infrastructure services at a relatively low cost from cloud computing [6]. Applications should be allocated enough resources to run smoothly. However, the application does not run at the highest load, preconfigured resources are idle, which would cause a waste of resources in most cases [7]. In addition, the pre-allocated resources may be not enough. In order to maximize the benefits of service providers while maintaining high quality of service (QoS), data centers need an efficient and dynamic resource expansion and allocation strategy [8]. Cloud computing usually uses specific prediction algorithms to predict resource demand and optimize resource allocation in order to improve resource utilization

✉ Yepeng Guan
ypguan@shu.edu.cn

[1] School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China

[2] Key Laboratory of Advanced Display and System Application, Ministry of Education, Shanghai 200072, China