Machine Learning Approaches for Load Balancing in Cloud Computing Services

Dalia Abdulkareem Shafiq
School of Computer Science & Engineering (SCE)
Taylor's University
Subang Jaya, Malaysia
daliakareem7@gmail.com

NZ Jhanjhi
School of Computer Science & Engineering (SCE)
Taylor's University
Subang Jaya, Malaysia
noorzaman.jhanjhi@taylors.edu.my

Azween Abdullah
School of Computer Science & Engineering (SCE)
Taylor's University
Subang Jaya, Malaysia
Azween.Abdullah@taylors.edu.my

2021 National Computing Colleges Conference (NCCC) | 978-1-7281-6719-0/20/$31.00 ©2021 IEEE | DOI: 10.1109/NCCC49330.2021.9428825

***Abstract*—As the demand for cloud services increases, optimization of resources becomes essential. Static algorithms are no longer sufficient to solve cloud-related challenges such as imbalanced workload distribution in Virtual Machines or improper resource allocation to cloud users. Thus, the need to explore other rich approaches can greatly improve cloud applications' performance and tackle the above challenges. This research investigates the latest Machine Learning approaches that can tackle the above challenges in cloud environment. A comparison of these approaches included highlighting their strengths and weaknesses to induce a research gap useful for upcoming researchers in the field.**

Keywords—Cloud Computing; Machine Learning; Regression; Classification; Virtualization; Optimization.

I. INTRODUCTION

Cloud Computing has been around since 1997 and the demand for cloud services is in the rise. The technology offers wide range of online services through multiple delivery models namely: Platform as Service (PaaS), Software as Service (SaaS) and Infrastructure as service (IaaS). Cloud users and organizations [1] mostly benefit from SaaS model whereby it can eliminate hardware cost by providing accessible software online such as Google Drive, Gmail, YouTube etc. This is one of the simplest delivery models in the cloud as it provides flexible services with less cost and physical storage. However, the increase in the services' everyday usage and handling the large data sets behind it can cause some challenges to the cloud services' backend.

IaaS model handles the backend (server-end) of the cloud services. Cloud computing relies heavily on virtualization to create an abstract layer between software and hardware [2][3]. In a typical cloud environment, users submit their requests, and these are converted into Virtual Machines. Cloud service providers in this model are responsible for ensuring efficient allocation of resources [4] to clients with optimal cloud services performance. Performance has been stated among the top three challenge in cloud computing [5], hence it becomes important to research efficient methods to improve the utilization of cloud resources and increase the satisfaction of users.

There has been extensive research on many objective-oriented algorithms to address optimization and resource allocation challenges in the cloud environment. However, the existing Load Balancing algorithms are either static where fewer parameters are considered or dynamic, where performance can easily degrade due to sudden failures in the nodes [6]. Thus, such approaches may not be suitable for use in in such environments where the load is constantly changing [7]. Therefore, it becomes important to discover other efficient and intelligent approaches to tackle cloud-related challenges.

Machine learning is a subcategory of artificial intelligence that has been an active topic in IT and intelligent systems. Data is stored in large quantities in the cloud machines, and it can be trained to make precise predictions and evaluations based on analysis to perform tasks more efficiently. It is the fastest-growing field nowadays. According to a survey done by RightScale in 2019, recent research shows that Machine Learning plays a vital role in Cloud computing. It represents a figure of 786 professionals (48% of respondents) among different expertise are considering using Machine learning services in the future [8]. Machine learning is being offered as a cloud service. Thus, the combination of such technologies can be established as architectures in the future to cover different layers from business workflow to software. Machine Learning approaches can be divided into two main groups [9]:

- **Supervised Learning**: this type of learning provides a precise classification of a labelled data sample with a defined output. This means the algorithm has a specific outcome that can be predicted from a set of independent variables. Examples of algorithms include Linear Regression, Support Vector Machine, Neural Networks and Naïve Bayes classifiers.
- **Unsupervised Learning**: unlike supervised learning, this type of learning does not provide a clear pattern in the dataset; the data sample are unlabelled. Hence a model is trained to have minimum errors when learning the categorization of such information. Such training can be used mainly to solve clustering scenarios such as classification—for example, K-Means clustering and Fuzzy Clustering algorithms.