Chapter 3
A Study on Distributed Machine Learning Techniques for Large-Scale Weather Forecasting

Balaji V.
https://orcid.org/0009-0005-9140-8486
Vellore Institute of Technology, Chennai, India

Sivagami M.
Vellore Institute of Technology, India

ABSTRACT

The weather data generated, processed, and collected by the sensor or shared by IoT devices and mobile devices has significantly increased weather data collection in daily life. The data generation speed also accelerated, and a vast amount of data has been collected and stored in distributed databases, improving weather forecasting. Still, the conventional processing method for massive data is distributed and centralized computing, and this chapter looks into how distributed machine learning techniques help as to increase the processing speed. Some distributed frameworks that play a significant role in massive data, like MapReduce, have been trained and tested to resolve various machine learning problems in a distributed environment. The aim of this chapter will provide different information about datasets, issues, platforms, and optimized approaches in a distributed environment. So, researchers can use and deploy new techniques in machine learning algorithms. It helps the researchers develop new strategies in distributed computing environments.

DOI: 10.4018/978-1-6684-9804-0.ch003

INTRODUCTION

The origins of human history: human have always aimed to predict and learn the world and strive to make better predictions. The weather is known as fluctuation in the daily atmosphere (Jain & Jain, 2017). Weather data like precipitation, temperature, pressure, wind speed, direction, humidity, air pressure, etc., collected from the different metrological departments from radar observation, sea-level observation, ground-level observation, and various data observations show the current weather status. Weather forecasting plays a vital role in many people's daily lives (Leu et al., 2014). It will impact many fields like marine trade, agriculture, aquaculture, industry, mining irrigation, etc.,

The enormous value of the dataset is utilized in various fields like geographical, industry, agriculture, and education (Thusoo et al., 2010). This chapter mainly concentrates on geographical data like NETCDF and Radar data. Parallel computing is complicated in the database and the cost is high in procedure to implement structured query language for developing the machine learning models. Data has categorized into three types there are structured (Standard format), semi-structured (Sensor data and Radar data), unstructured data (multimedia data), and reason for data growth by the database parallelization(Al-kahtani & Karim, 2017).

Machine learning approaches are applied in various applications and areas, such as robotics, natural language processing, clustering(Yeo et al., n.d.), and document classification(Shahid et al., 2019). Massive data like terabytes are applying the machine learning model increased drastically and the cost is also high for optimization by parallel computing. Distributed computing is concrete in the machine learning model; it will divide the workload among the serval computational machine to complete the work(Neumeyer et al., 2010).

Distributed Systems

Distributed systems are a combination of various nodes or machines that must have computational capabilities and communicate with each other, as shown in Fig 2. The data transfer between the machine is performed through the network with efficient bandwidth with N number of resources. In a distributed system, the single node failure cannot affect a network's overall communication and reliability; it will improve data transmission and easy access with nodes at the maximum workload. The distributed systems must ensure minimum communication cost with nodes and utilize the system's computational capability.

Generally, MapReduce and Hadoop approaches are exploited to implement the large-scale model in machine learning algorithms. The Hadoop system is most