An approximation method of regression analysis in concurrent big data stream

Chanintorn Jittawiriyanukoon, author

Abstrak

Time series big data dynamically changes the size, and, unfortunately, it may be difficult to curate the enormous amount of data due to the processing capacity and storage size. This big data allows researcher to iterate on the model millions of times over. To execute a regression on several billion rows of data on a distributed network, the resource capacity regarding large volumes of data and its distributed environment must be considered. Algorithms must be real-time based data awareness. Moreover, analyzing big data sources requires the data to be pre-processed rather than immediately collected and analyzed. This pre-processing approach for the big data sources helps minimize the amount of collected data by extracting insights. It analyzes big data quicker and is cost-effective for storage space. Hence, in this research, an approximation method for analyzing regression problems in a big data stream with parallelism is proposed. The partitioning method for huge data stream helps reduce the computing time and required space, and the speed-up can improve the processing time. The performance evaluation of concurrent regression model is first executed by massive online analysis (MOA) simulation. Then, to validate the approximation method, the results performed by our proposed method are compared to those results collected from the simulation. The comparisons show evenly between the two methods.