World Academy of Science, Engineering and Technology
International Journal of Computer and Systems Engineering
Vol:14, No:07, 2020

An Improved Parallel Algorithm of Decision Tree

Authors : Jiameng Wang, Yunfei Yin, Xiyu Deng

Abstract : Parallel optimization is one of the important research topics of data mining at this stage. Taking Classification and Regression Tree (CART) parallelization as an example, this paper proposes a parallel data mining algorithm based on SSP-OGini-PCCP. Aiming at the problem of choosing the best CART segmentation point, this paper designs an S-SP model without data association; and in order to calculate the Gini index efficiently, a parallel OGini calculation method is designed. In addition, in order to improve the efficiency of the pruning algorithm, a synchronous PCCP pruning strategy is proposed in this paper. In this paper, the optimal segmentation calculation, Gini index calculation, and pruning algorithm are studied in depth. These are important components of parallel data mining. By constructing a distributed cluster simulation system based on SPARK, data mining methods based on SSP-OGini-PCCP are tested. Experimental results show that this method can increase the search efficiency of the best segmentation point by an average of 89%, increase the search efficiency of the Gini segmentation index by 3853%, and increase the pruning efficiency by 146% on average; and as the size of the data set increases, the performance of the algorithm remains stable, which meets the requirements of contemporary massive data processing.

Keywords : classification, Gini index, parallel data mining, pruning ahead
Conference Title : ICDM 2020 : International Conference on Data Mining
Conference Location : Dublin, Ireland
Conference Dates : July 30-31, 2020