Efficient Multi-Sample Approximate Computing for Scalable Analysis of Massive Distributed Datasets on Resource-Constrained Clusters.

Alladoumbaye Ngueilbaye[1,2]*, Joshua Zhexue Huang [1, 2, 3], Yongda Cai [1,2], Xudong Sun [4]

[1]National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University,518060, China
[2] Big Data Institute, College of Computer Science and Software Engineering, Shenzhen University, 518060, China
[3]Guangdong Laboratory of Artificial Intelligence and Digital Economy, Shenzhen, 518107, China
[4]College of Management, Shenzhen University, Shenzhen, 518060, China
angueilbaye@szu.edu.cn, zx.huang@szu.edu.cn, caiyongda2021@email.szu.edu.cn, sunxudong2016@email.szu.edu.cn

Abstract

The prolific explosion of data in today's digital sphere by modern AI applications has created new challenges and opportunities for business industries. This has necessitated the development of scalable methods for analyzing massive datasets stored in distributed systems. However, resource-constrained clusters often struggle to process such datasets due to memory constraints and the computational overhead of distributed AI algorithms. This paper proposes efficient multi-sample approximate computing (EMSAC), a novel approach designed to enable scalable analysis of massive distributed datasets on small clusters with limited memory. EMSAC leverages multiple small random samples, processed in parallel using sequential algorithms, to approximate the analysis of the entire dataset. The approach has been implemented in Spark using the LOGO computing framework to address three key challenges: (1) efficient generation of multiple small random samples from a massive distributed dataset; (2) conversion of these data block samples to a partial RSP data model and parallel execution of sequential algorithms on the partial RSP data model to mine frequent itemsets; and (3) aggregation of each data block result to produce the approximate set of frequent itemsets of $\mathbb{D}$. To guarantee the quality of random data block samples, we theoretically provide a bound on the estimated number of data blocks to be selected from the distributed data file. Empirical evaluations on synthetic and real-world datasets demonstrate that EMSAC outperforms traditional distributed and sampling-based approaches in terms of scalability, accuracy, and computational efficiency. The findings have shown that EMSAC is suitable for processing massive distributed data and generating accurate approximate frequent itemsets with constrained clusters.

Acknowledgments

This research has been supported by the National Engineering Laboratory for Big Data System Computing Technology under Grant No. SZU-BDSC-IF2024-03 and the National Natural Science Foundation of China under Grant No. 61972261.

*Corresponding author
Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.