Tile Low-Rank Approximation of Maximum Large-Scale Likelihood Estimation on Manycore Architectures

ExaGeoStat: A Framework for Large-Scale Weather and Climate Prediction using Machine Learning

SAMEH ABDULAH, HATEM LTAIEF, YING SUN, MARC GENTON AND DAVID KEYES
EXTREME COMPUTING RESEARCH CENTER, KING ABDULLAH UNIVERSITY OF SCIENCE AND TECHNOLOGY.
Motivation

• **Maximum Likelihood Function**
  • An important statistical technique for predicting unknown measurements in climate and environmental applications

• **Weather and climate data are usually big!**
  • Prohibitive computational Cost and memory requirements
  • Two real examples: Soil Moisture and Wind Speed

• **Exascale GeoStatistics (ExaGeoStat)**
  • A framework which exploits machine learning, statistical modeling and forecasting, and the state-of-the-art linear algebra techniques to handle large-scale geostatistics data
ExaGeoStat

Machine Learning
- Maximum Likelihood Estimation (MLE).
- Supervised Learning.

Big Data

Linear Algebra
- Dense Computation.
- Tile Low-Rank Approximation
ExaGeoStat Components

• **Synthetic Dataset Generator**
  • Generate large-scale geospatial datasets which can be separately used as benchmark datasets for other software packages

• **Maximum Likelihood Estimator (MLE)**
  • Evaluate the maximum likelihood function on large-scale geospatial datasets
  • Support full machine precision accuracy (full-matrix) and Tile Low-Rank (TLR) approximation

• **ExaGeoStat Predictor**
  • Predict unknown measurements on known geospatial locations by leveraging the MLE estimated parameters
Outcomes

• Hardware-agnostic framework
• 2M spatial locations (160 TB memory requirement)
• 96 % prediction accuracy on real datasets
• Large synthetic spatial data generation tool
• R-Wrapper package
ExaGeoStat Under the Microscope

• ExaGeoStat is an open-source software which is available at [https://github.com/ecrc/exageostat](https://github.com/ecrc/exageostat)

• ExaGeoStat 0.1.0 (Nov. 9th 2017)
  • Support exact Computation using Chameleon dense Linear algebra library and StarPU runtime system
  • Support real and synthetic geospatial datasets
  • Soil moisture dataset at Mississippi basin area

• Today:
  • ExaGeoStat supports
    • Tile-Low Rank approximation (TLR) using HiCMA TLR approximation library and StarPU runtime system
    • Performance results of TLR-based computations on shared and distributed-memory systems attain up to 13X and 5X speedups
    • Support NetCDF Format
    • Win Speed dataset at Middle-East
**Title:** Low-Rank Approximation of Maximum Large-scale Likelihood Estimation on Manycore Architectures

**Authors:** Saud Albarshi, Haitham Leblebici, Ying Chen, Marc Canton, and David Keys

**Abstract:**

This paper presents a framework for approximating the maximum likelihood function for large-scale spatial data. The framework, called ExGeoStat, is designed to handle the computational challenges posed by manycore architectures. It utilizes low-rank approximation techniques to efficiently process large datasets.

**Dataset Generator:**

- Generates synthetic spatial data by randomly sampling points from a specified distribution.
- Supports both synthetic and real-world datasets.
- Produces data in formats that are compatible with manycore architectures.

**ExGeoStat Predictor:**

- The associated conditional distribution where $Z$ represents a known parameter set and $\Theta$ represents the unknown parameter set.
- Utilizes the low-rank approximation to efficiently compute the likelihood function.

**Synthetic Data Accuracy Verification:**

- Measures the accuracy of the predicted results against known synthetic datasets.
- Validates the framework's ability to handle large-scale spatial data.

**Real Data Accuracy Evaluation:**

- Evaluates the performance of the framework on real-world datasets.
- Compares the predicted results with actual data to assess accuracy.

**Software Release:**

- The ExGeoStat package is available for download at [GitHub](https://github.com/ExGeoStat).
- The package includes detailed documentation and examples for easy integration.

**Performance on Shared Memory:**

- Evaluated on two socket 16-core Intel Xeon Gold 6240.
- Showcases the performance gains achieved with the ExGeoStat framework.

**Performance on Distributed Memory:**

- Evaluated on a cluster of 256 nodes.
- Demonstrates the scalability and efficiency of the framework in a distributed environment.