Study on Distributed and Parallel Non-Linear Optimization Algorithm for Ocean Color Remote Sensing Data

Jung-Ho Um, Sunggeun Han, Hyunwoo Kim, Kyongseok Park
Scientific Data Technology Lab
Korea Institute of Science and Technology Information
Daejeon, Rep. of Korea
{jhum, sghan, pardess, gspark}@kisti.re.kr

***Abstract*— recent developments in science and technology have made it possible to analyze data observed by satellites using optical properties. By monitoring changes in the ocean environment and ecosystem, we are currently conducting ocean environmental studies to identify abnormal weather phenomena. International aerospace laboratories such as NASA and ESA are publishing these observed data to ocean scientists around the world. Satellite sensing data accumulates day by day, but data volume for the global scale is so large that scientists usually divide the space for only the area of interest and perform time series analyses. Time series analysis is mainly applied to nonlinear distributions. However, studies of the ocean environment require analysis of the global ocean and ocean ecosystems. Data analysis in the global domain requires non-linear data fitting for every cell of the satellite imagery data. However, commercial and open-source data analysis tools such as Matlab or R do not provide non-linear data fitting for multiple cells. Because of this, there is a difficulty for ocean scientists to directly implement the analysis of data and it is hard to guarantee distributed and parallelized computation performance. Therefore, in this paper, we propose an algorithm that can distribute and parallelize, in a multi-dimensional database environment, the Levenberg-Marquadt (LM) algorithm, which is well known as a non-linear data fitting algorithm. Our algorithm achieved about 7.5 times speed–up on average, compared to the MINPACK LM algorithm, which is based on MPI and written in FORTRAN. In addition, our algorithm improved 74.3 times speed-up when comparing to the maximum performance for each algorithm. As future research, we will utilize the developed algorithms in the ocean science field for data analysis of global scale satellite imagery data.**

Keywords—distribute and parallelized algorithm, non-linear data fitting, remote sensing data, ocean science

I. INTRODUCTION

With advances in aerospace technology, research on large-scale satellite imagery data analysis is being conducted by international aerospace laboratories such as NASA and ESA [1, 2]. In studying ocean ecosystems and environmental changes, the optical properties in the remote sensing data are analyzed according to the light wavelength. This enables an analysis of the density of chlorophyll, the primary producer of photosynthesis using light in ocean color sensing data. Through this analysis, by observing how ocean environment and ecosystem change, we can produce analytical information to solve ocean environmental problems such as abnormal climate or red tide. SeaWIFS [3] and MODIS [4] which has been in operation since 1997 and 2001, respectively, are mainly used for this research. Based on these large-scale satellite imagery data, researchers who analyze remote sensing data hope to analyze various living ocean organisms around the globe. However, the data volume is very large, and software and hardware are not sufficient, making data difficult to analyze. Therefore, in this paper, we develop distributed / parallel non-linear data fitting algorithms such as the Gaussian fitting algorithm, which is frequently used in ocean biological distribution analysis. For this, we design a distributed / parallel algorithm for the well-known Levenberg-Marquadt (LM) algorithm [5]. In addition, for analysis of satellite imagery data from a spatiotemporal aspect, as well as data processing and management, we implemented a distributed / parallel method using the SciDB plugin, which is a multidimensional and parallel database. The contribution of this paper is as follows:

- The proposed algorithm can be concurrently fitted to 3D satellite images by distributing / parallelizing the Levenberg-Marquadt algorithm, which performs curve fitting on two-dimensional data.
- By minimizing disk I / O in chunk units and concurrent data computation for non-linear data fitting, the distributed and parallelized LM algorithm achieve better performance gain than that of the well-known FORTRAN-based MINPACK algorithm.
- We tried to find an optimized chunk size for the parallelism performance that SciDB can provide by experimenting with changing the levels of scatter / parallelism based on the SciDB chunk.

II. DISTRIBUTED AND PARALLEL NON-LINEAR OPTIMIZATION

A. Non-Linear Optimization Use Case in Ocean color study

In this paper, we develop a distributed and parallel non-linear curve fitting algorithm to understand the trend of ocean suspended chlorophyll distribution. To do this, we analyze the requirements for calculating chlorophyll distribution trends. To calculate the density of ocean suspended chlorophyll, we use a non-linear fitting that is well suited to the Gaussian data

978-1-5386-3472-1/17/$31.00 ©2017 IEEE

distribution, as shown in Figure 1. This is because there is a period every six months when chlorophyll flourishes [6, 7].The fitting of the curve starts with the initial values for each of the following parameters:

A<- Maximum value of Y axis

B<- value of x axis when y=A

C<- 8 (Minimal value considering time duration and chlorophyll boom period)

D<- mean value of Y values – standard deviation of Y values

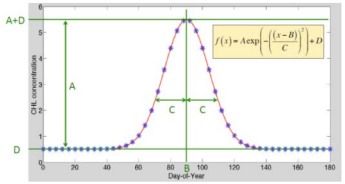

Figure 1. Curve fitting for chlorophyll density

B. Algorithm

The existing LM algorithm functions to fit two-dimensional data to nonlinear functions as shown in Figure 1. However, in order to concurrently compute the chlorophyll density (concentration) of a 4320 * 8640 size cell over time based on the L3 of the NASA MODIS remote sensing data, the LM algorithm must be processed in parallel for each cell. The proposed algorithm performs the LM algorithm concurrently through distributed / parallel operation for each chunk defined by SciDB. Because the parameters of all the cells in each chunk are kept in vector format, the parameter values estimated through curve fitting in multiple cells are preserved during curve fitting while traversing other cells. This allows non-linear curve fitting of marine suspended bio-fluctuation patterns over time in all cells, covering the global area. Our algorithm is presented in Algorithm 1 based on the above design considerations.

Algorithm 1. Distributed and Parallel LM Algorithm

Input: 3-D Array consisting of Cell(i,j,t) i∈Latitude, j∈Longitude, t∈Time

Output: β ={A(i,j),B(i,j),C(i,j),D(i,j) i∈Latitude, j∈Longitude}

1. A(i,j)<-max(cell(i,j,t=1), cell(i,j,t=2), … cell(i,j,t=N)
2. B(i,j)<- time t of cell A(i,j), C(i,j)<-Constant value
3. D(i,j)<- avg($\sum_{t=1}^{N}$ cell(i, j, t))-stdev($\sum_{t=1}^{N}$ cell(i, j, t))
4. δ_{prev} <- 0
5. while $\frac{S(\beta)-S(\beta+\delta)}{S(\beta)}$ > Threshold
6. for each chunk in parallel do
7. for each cell(i,j)
8. J<-Gradients(A(i,j),B(i,j),C(i,j),D(i,j))
9. δ<- LU-decompose($J^TJ+(1+\lambda)$diag(J^TJ), J^T[Y- F(β)])
10. δ_2<- LU-decompose($J^TJ+\frac{(1+\lambda)}{v}$diag($J^TJ$), J^T[Y- F(β)])
11. If((RMSE(δ)), RMSE(δ_2)) > RMSE(δ_{prev}) then
12. $\beta = \beta + \delta * v$
13. Else If(RMSE(δ) > RMSE(δ_2)) then
14. $\beta = \beta + \delta_2$, $\delta_{prev}=\delta_2$
15. Else $\beta = \beta + \delta$, $\delta_{prev}= \delta$

III. Preliminary Evaluation Results

We compare the proposed algorithm and the MPI - based MinPack LM algorithm in terms of execution time. For the experimental environment, we use a cluster consisting of 7 nodes with an Intel Xeon CPU E7-4860* 40 cores CPU, with 256GB Memory, and a 10TB HDD. The data were clipped from the MODIS Aqua L3 data centered on the coast of the Korean peninsula (4 km, 471 × 271). Threshold is set to 10^{-6}. In order to specify the same number of processes in SciDB as in MPI, experiments were performed by mapping the chunk size, a parallel unit, to the number of processes in the corresponding MPI. As a result, the proposed method has an average performance of 7.5 times speed-up. As the number of MINPACK process increases, more time consumed due to the communication overhead. The execution time of our algorithm is reduced to 1.35 seconds when chunk size is 10. Therefore, the performance can be improved to 75 times comparing to the maximum performance of MINPACK of 100.37 seconds.

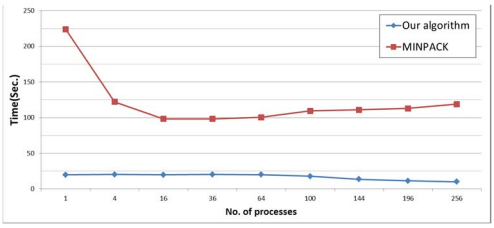

Figure 2. LM algorithm execution time comparison

IV. Conclusion

In this paper, we present a distributed and parallelized Levenberg-Maquadt algorithm that is often used in marine science research on SciDB. The proposed algorithm showed better performance than the existing MINPACK. As Future research, we will carry out data analysis on global image data of various resolutions.

Acknowledgement

This work was supported by a grant (K-17-L03-C01-S03) funded by the ministry of science and ICT, Korea.

References

[1] NASA Climate Change, https://climate.nasa.gov/

[2] ESA Climate Change Initiative, http://cci.esa.int/

[3] NASA SeaWIFS, https://oceancolor.gsfc.nasa.gov/SeaWiFS/

[4] NASA MODIS, https://modis.gsfc.nasa.gov/

[5] Marquardt, Donald (1963). "An Algorithm for Least-Squares Estimation of Nonlinear Parameters". SIAM Journal on Applied Mathematics. 11 (2), 1963, pp 431–441.

[6] Kim, H.C., S. Yoo, and I.S. Oh, "Relationship between phytoplankton bloom and wind stress in the sub-polar frontal area of the Japan/East Sea", J. Mar.Sys.67, 2007, pp.205-216.