한국정보전자통신기술학회논문지(jkiiect)'23-12, Vol.16 No.6
논문 23-16-06-489 http://dx.doi.org/10.17661/jkiiect.2023.16.6.489

Low-GloSea6 기상 예측 모델 기반의 비선형 회귀 기법 적용 연구

박혜성, 조예린, 신대영, 윤은옥, 정성욱*

A Study on Applying the Nonlinear Regression Schemes to the Low-GloSea6 Weather Prediction Model

Hye-Sung Park, Ye-Rin Cho, Dae-Yeong Shin, Eun-Ok Yun, Sung-Wook Chung*

요 약 하드웨어의 성능 및 컴퓨팅 기술의 발전 덕분에 기후환경 변화를 대비하기 위해 기후예측 모델 또한 발전하고 있다. 한국 기상청은 GloSea6를 도입하여 슈퍼컴퓨터를 이용하여 기상 예측을 하고있으며, 각 대학 및 연구 기관에서는 중소규모 서버에서 사용하기 위해 저해상도 결합모델인 Low-GloSea6를 사용하여 기상 연구에 활용하고 있다. 본 논문에서는 중소규모 서버에서의 기상 연구의 원활한 연구를 위해 Low-GloSea6의 Intel VTune Profiler를 사용한 분석을 진행하였으며 1125.987초의 CPU Time을 수행하는 대기모델의 tri_sor_dp_dp 함수를 Hotspot으로 검출하였다. 수치적 연산을 진행하는 기존 함수에 머신러닝 기법의 하나인 비선형 회귀모델을 적용 및 비교하여 머신러닝 적용 가능성을 확인하였다. 기존 tri_sor_dp_dp 함수의 실제 연산되는 값인 1e-3 ~ 1e-20의 범위를 가지는 Output Data인 변수 "Px"를 기준으로 평가하였을때 K-최근접 이웃 회귀 모델은 MAE가 1.3637e-08, SMAPE가 123.2707%로 가장 우수하게 나타났으며 RMSE의 경우 Light Gradient Boosting Machine 회귀 모델이 2.8453e-08로 가장 우수한 성능을 보이는 것으로 측정되었다. 따라서 Low-GloSea6 수행 과정 중 tri_sor_dp_dp 함수의 데이터를 추출 후 비선형 회귀 모델을 적용한 결과로 기존의 tri_sor_dp_dp 함수의 수치적 연산 값과 K-최근접 이웃 회귀 모델을 비교하였을 때 SMAPE가 123.2707%의 오차가 발생하는 것으로 측정되어 기존 모듈의 대체 가능성이 있다는 것을 확인하였다.

Abstract Advancements in hardware performance and computing technology have facilitated the progress of climate prediction models to address climate change. The Korea Meteorological Administration employs the GloSea6 model with supercomputer technology for operational use. Various universities and research institutions utilize the Low-GloSea6 model, a low-resolution coupled model, on small to medium-scale servers for weather research. This paper presents an analysis using Intel VTune Profiler on Low-GloSea6 to facilitate smooth weather research on small to medium-scale servers. The tri_sor_dp_dp function of the atmospheric model, taking 1125.987 seconds of CPU time, is identified as a hotspot. Nonlinear regression models, a machine learning technique, are applied and compared to existing functions conducting numerical operations. The K-Nearest Neighbors regression model exhibits superior performance with MAE of 1.3637e-08 and SMAPE of 123.2707%. Additionally, the Light Gradient Boosting Machine regression model demonstrates the best performance with an RMSE of 2.8453e-08. Therefore, it is confirmed that applying a nonlinear regression model to the tri_sor_dp_dp function during the execution of Low-GloSea6 could be a viable alternative.

Key Words : Global Climate Model, Low-GloSea6, UM, Machine Learning, Nonlinear Regression

1. 서론

고도화된 컴퓨팅 기술의 발전 및 하드웨어의 성능이 향상됨에 따라 기후환경 변화를 대비하기 위해 기후 예측 모델 또한 기술이 발전되고 있다. 또한, 일상생활 및 산업환경에서의 기후예측의 중요성이 확대됨

This research is funded by the Mid-level professor Financial Program at Changwon National University in 2023
*Corresponding Author : Department of Computer Engineering, Changwon National University (swchung@changwon.ac.kr)
Received December 10, 2023 Revised December 12, 2023 Accepted December 14, 2023

으로써 기후예측의 정밀도 및 실시간 예측의 필요성이 증가하고 있다.

수치적 기상 예보 모델 중 하나인 지구 기상 예보 모델(Global Climate Model, GCM)은 전 지구적 기상 분석 및 예측하는 모델이다[1]. 한국 기상청(KMA, Korea Meteorological Administration)에서도 슈퍼컴퓨터 5호기를 도입하여 기상 예측분야에서 많은 연구가 이루어지고 있으며 더욱 정밀한 기상 연구를 위해 영국 기상청으로부터 2014년에 GloSea5(Global Seasonal Forecasting System 5)을 도입하였다[2]. 그 이후 한국 기상청은 GloSea5의 과학적 구성(Science Configuration)을 GA3 (Global Atmosphere version 3; 전구 대기모델의 구성 버전)에서 GC2 (Global Coupled version 2; 결합모델의 구성 버전)으로 업그레이드하였고, 2021년 결합모델 구성을 GC2에서 GC3.2로 업그레이드한 GloSea6를 설치하여 기상청 현업 기후예측시스템으로 운영 중에 있다[3, 4].

슈퍼컴퓨터를 사용하지 못하는 각 연구기관, 기업 및 대학 연구소에서는 중소규모 서버를 기반으로 한 실험환경에서 기존의 GloSea6를 직접 구동하기에는 하드웨어 스펙의 한계가 존재한다. 따라서 격자 해상도를 낮춘 저해상도 결합모델인 Low-GloSea6로 대체하여 사용 중이며 본 논문에서도 저해상도 결합모델을 사용하였다[5].

중소규모 서버에서 높은 사양의 하드웨어가 필요한 기상예측 모델을 수행하기 위한 또 다른 방법으로는 최근 많은 발전이 이루어진 머신러닝 기법을 적용하여 시간적 단축을 이루어 기존 수치적 예보 모델과 비교하였을 때 상대적으로 낮은 사양에서도 수행할 수 있다는 장점이 있다[5, 6].

이를 통해 본 논문에서는 중소규모 서버에서 활발한 기상 연구를 위해 전 지구적 기상 예측 모델의 저해상도 결합모델 버전인 Low-GloSea6를 사용하여 VTune Profiler를 사용하여 분석 진행 후 가장 오랜 시간 소요된 Hotspot을 검출하여 해당 Hotspot에서 실제 파라미터 및 변수로 사용되는 데이터를 추출한다. 그런 다음 머신러닝 기법의 하나인 비선형 회귀(Nonlinear Regression)에 데이터 기반 학습을 진행하여 기존 수치적 연산 방식과 비선형 회귀 모델을 평가 및 비교 진행 후 Low-GloSea6 수행과정 중 수행시간 단축을 위한 머신러닝 적용 가능성을 입증한다.

2. Low-GloSea6

2.1 Low-GloSea6 구성

GloSea6는 전 지구 기후모델로 HadGEM3 (Hadley Center Global Environmental Model Version 3)를 기반으로 개발된 앙상블 예측모델이며, 해상도는 대기 및 지면은 N216, 해양 및 해빙은 eORCA025를 사용한다. 연직 해상도는 대기의 경우 85층, 해양의 경우 75층이다. GloSea6는 표 1과 같이 대기, 지면, 해양, 해빙 총 4가지의 모델이 서로 결합되어 구성된다. Low-GloSea6는 GloSea6와 같은 구조를 이루고 있으며 대기, 지면, 해양, 해빙을 구성하고 있는 각 격자의 해상도를 낮춘 저해상도 결합모델이다[4].

표 1. GloSea6 모델
Table 1. Model of GloSea6

model	name	purpose
UM	Met Office Unified Model	Atmosphere
CICE	Los Alamos Sea Ice Model	Sea Ice
NEMO	Nuleus for European Modeling of the Ocean	Ocean
JULES	Joint UK Land Environment Simulator	Land

각 4가지 모델은 초기자료들을 이용하여 수치모델 적분을 진행한다. 대기 및 지면 모델은 15분, 해양 및 해빙모델은 20분 간격으로 적분을 진행한다. 이때 GloSea6 기준으로 대기 및 지면 모델은 60km의 해상도이지만 해양 및 해빙 모델은 25km의 해상도로 격자의 크기가 달라 OASIS 결합자를 통해 1시간 간격으로 변수를 교환하여 4가지 모델이 유기적으로 수행되도록 한다. OASIS 결합자로는 OASIS3-MCT를 사용하여 sparse matrix 연산법을 사용하는 성분 모델

프로세스 간 직접 교환을 진행한다[7]. 대기 및 지면 모델의 산출물은 ff(filed file) 파일 형식으로, 해양 및 해빙 모델의 산출물은 netCDF 파일 형식으로 산출된다[3].

2.2 Low-GloSea6 프로파일링

본 논문에서는 Low-GloSea6의 Hotspot 검출을 위하여 Intel에서 개발 및 제공하는 VTune Profiler를 사용하여 분석을 진행하였다. Vtune Profiler는 어플리케이션 성능, 시스템 성능 및 최적화를 위한 도구를 제공한다. 프로파일링을 통해 어플리캐이션 동작 시 코드 내에서 가장 오랜 시간 소요되는 Hotspot을 탐색할 수 있다. 그림 1과 같이 VTune Profiler를 사용하여 Cluster, Node, Core에 맞게 MPI, CPU, Memory, Thread-level, FPU Bound 및 이슈 성능과 튜닝이 가능하다[8].

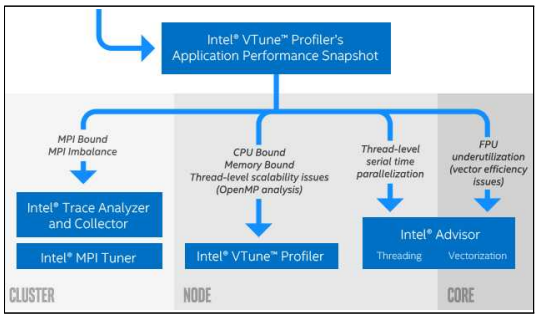

그림 1. VTune 프로파일러[9]
Fig. 1. VTune Profiler

표 2. 프로파일링에 수행된 시스템 스펙
Table 2. System specs used for profiling

Name	Hardware Specification
CPU	Intel® Core 10th GEN i7-10700K
RAM	64GB
SSD	1TB
GPU	GeForce RTX 3080

본 논문에서는 CPU Bound 분석을 위해 그림 1의 시나리오 중 Node 기반 Intel Vtune Profiler를 사용하였고 CPU Usage 기반 Hotspot 분석을 시도하였다. VTune Profiler는 두 가지 샘플링 기반 탐색 모드가 존재하는데 오버헤드가 존재하지만, 수집을 위해 샘플링 드라이버가 필요하지 않은 사용자 모드 샘플링과, 최소 수집 오버헤드를 샘플링 드라이버를 설치해야 하는 하드웨어 이벤트 기반 샘플링이 존재한다. 본 논문에서는 샘플링 드라이버를 설치하지 않는 사용자 모드를 사용하여 샘플링 Hotspot 분석을 시도하였다. 프로파일링을 진행한 서버의 사양은 표 2와 같다.

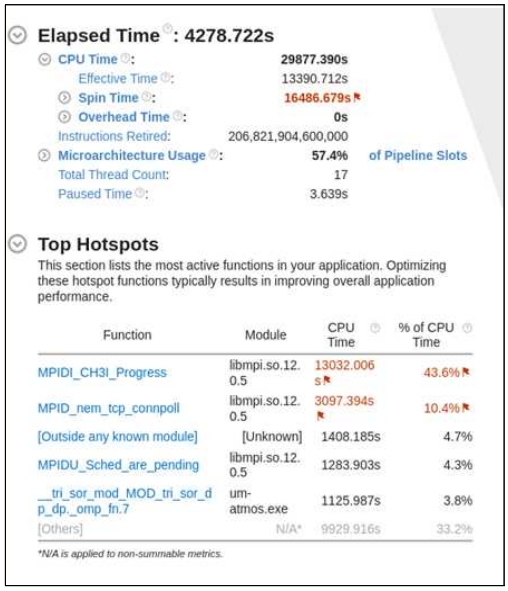

그림 2. Low-GloSea6 분석 결과
Fig. 2. Low-GloSea6 Profiling Result

Intel VTune Profiler를 사용하여 Low-GloSea6를 분석한 결과 그림 2와 같이 1일 차 기준 Elapsed Time은 4278.722초, 전체 CPU Time은 29877.390초 소요되었다. 그중 Effective Time은 13390.712초이며, Spin Time은 16486.679초로 절반 이상의 시간으로 Spin Time이 비중을 차지하고 있다. 그림 3과 같이 Low-GloSea6 수행 중 모듈별 CPU Time을 비교하였을 때 가장 오래 수행하는 모듈로는 GloSea6의 4가지 모델 중 대기(UM) 모델의 tri_sor.F90 모듈의 tri_sor_dp_dp 함수가 가장 많은 CPU Time이 소요된 Hotspot으로 검출되었다. 그림 3의 5번째 행에서 볼수 있듯이 해당 함수는

1125.987초 소요되었으며 13.3%

CPU Time ▼	Instructions Retired	Microarchitecture Usage	Module	Function (Full)
13032.006s	120,687,327,400,000	79.8%	libmpi.so.12.0.5	MPIDI_CH3I_Progress
3097.394s	21,479,359,400,000	36.6%	libmpi.so.12.0.5	MPID_nem_tcp_connpoll
1408.185s	4,498,531,200,000	34.6%		[Outside any known module]
1283.903s	17,827,578,400,000	100.0%	libmpi.so.12.0.5	MPIDU_Sched_are_pending
1125.987s	1,706,568,600,000	13.3%	um-atmos.exe	__tri_sor_mod_MOD_tri_sor_dp_dp._omp_fn.7
888.238s	4,721,956,000,000	48.4%	libm-2.17.so	__ieee754_pow_sse2
617.219s	4,255,281,800,000	63.0%	libm-2.17.so	__ieee754_exp_avx
450.796s	2,156,222,600,000	41.3%	libm-2.17.so	__exp1
437.765s	212,013,400,000	23.0%	libgfortran.so.4.0.0	func@0x1c270
347.046s	986,077,200,000	22.4%	um-atmos.exe	__mod_cosp_MOD_cosp_iter
334.226s	1,264,324,600,000	43.5%	libmpi.so.12.0.5	MPID_nem_network_poll
315.775s	2,268,923,000,000	64.5%	um-atmos.exe	__eg_cubic_lagrange_mod_MOD_eg_cubic_lagrange._
315.586s	1,213,792,200,000	43.6%	libmpi.so.12.0.5	func@0x3da00
241.898s	245,096,200,000	8.3%	um-atmos.exe	__glue_conv_6a_mod_MOD_glue_conv_6a
168.967s	1,118,036,000,000	49.7%	libc-2.17.so	_int_malloc
143.053s	496,743,600,000	29.9%	um-atmos.exe	__eg_calc_ax_mod_MOD_eg_calc_ax_dp_dp._omp_fn.
137.651s	868,364,600,000	57.4%	um-atmos.exe	bdy_impl3_._omp_fn.0
128.050s	123,070,600,000	14.9%	libgfortran.so.4.0.0	func@0x1c230
115.389s	53,329,200,000	18.5%	libgfortran.so.4.0.0	func@0x1c4b0
115.055s	849,501,400,000	54.2%	libc-2.17.so	_int_free
114.831s	294,956,000,000	16.7%	libmpi.so.12.0.5	MPIDI_CH3U_Recvq_FU
93.374s	176,665,800,000	15.9%	um-atmos.exe	__ls_ppnc_mod_MOD_ls_ppnc_gather

그림 3. Low-GloSea6 모듈별 분석 결과
Fig. 3. Profiling Results for each Low-GloSea6 module

Microarchitecture Usage로 여러 가지 모델에서 동시에 호출하여 다양한 값을 병렬적으로 연산하기 때문에 비교적 낮은 수치로 분석되었다.

2.3 tri_sor.F90

앞서 Low-GloSea6를 모듈별 수행시간을 분석하였을 때 가장 오랫동안 CPU Time을 소모하는 Hotspot으로 검출된 tri_sor.F90 모듈은 Low-GloSea6를 구성하는 4가지 모델 중 대기(UM) 모델에 존재하며 dynamics solver 기능을 하는 모듈이다. Low-GloSea6 연산 과정에서 선형 시스템의 해를 근사적으로 계산하고 병렬화를 개선하는 역할을 하며 tri_sor.F90 모듈 내에는 tri_sor_dp_dp, tri_sor_dp_sp, tri_sor_sp_dp, tri_sor_sp_sp 총 4가지 서브 루틴이 존재한다. 이 중 프로파일링을 통해 Hotspot으로 검출된 함수는 tri_sor_dp_dp 함수이며, 표 3과 같이 해당 함수에서 사용되는 파라미터로 Input data 9가지와 Output data 1가지를 사용한다.

표 3. tri_sor_dp_dp 변수
Table 3. tri_sor_dp_dp Variable

Input_Data	Output_Data
x	Px
ltHlm_Ln	
ltHlm_Ls	
ltHlm_Le	
ltHlm_Lw	
ltHlm_Ld	
ltHlm_Lu	
ltHu_k	
ltHd_k	

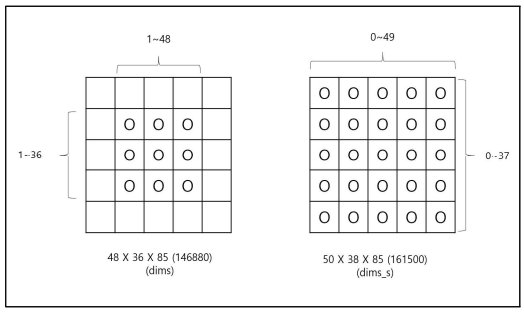

그림 4. tri_sor_dp_dp 격자 크기
Fig. 4. Grid Size of tri_sor_dp_dp

tri_sor_dp_dp 함수에서 사용되는 변수는 각자 다

른 격자 크기를 가지고 있다. 그림 4와 같이 변수들의 격자는 대기·지면 모델의 격자 크기(N216)를 기준으로 구성되어 있다. 대기·지면 모델의 격자와 동일한 크기인 dims_s(50*38*85) 격자와, dims_s에서 연직 높이는 동일하며 수평 및 수직 격자의 최외곽 부분을 1칸 축소한 크기인 dims(48*36*85)로 구성되어있다. 앞서 표 3에서 확인한 변수들 중 x와 Px 큰 격자인 dims_s이며, 나머지 8개의 변수는 작은 격자인 dims로 동작한다. 따라서 다른 크기의 격자들을 머신러닝에 적용하기 위하여 작은 크기 격자인 dims의 크기의 데이터를 사용하도록 전처리 후 비선형 회귀 모델을 적용하여 가능성을 확인하였다.

3. 비선형 회귀

그림 5와 같이 tri_sor_dp_dp 함수에서 사용되는 변수들의 히트맵을 작성하여 분석하였을 때 Output feature인 변수 “Px”와 나머지 변수와의 상관계수가 0에 가까운 값으로 측정되었다. 이는 상관관계가 거의 없는 것을 뜻하며 선형적인 관계가 아닌 비선형적인 관계를 의미한다. 따라서 본 논문에서는 tri_sor_dp_dp 함수를 머신러닝 학습하기 위하여 비선형 회귀(Nonlinear Regression) 기법을 적용하였다. 비선형 회귀란 독립변수와 종속변수 간 관계가 선형적이지 않을 때 적합하며 튜닝파라미터의 경우의 수가 다양해 모델의 유연성이 높다는 장점이 있다. 본 논문에서는 다층 퍼셉트론(Multilayer Perceptron, MLP), K-최근접 이웃(K-Nearest Neighbor, KNN), Light Gradient Boosting Machine(LightGBM) 3가지 회귀 모델을 구성하여 평가 및 비교한다.

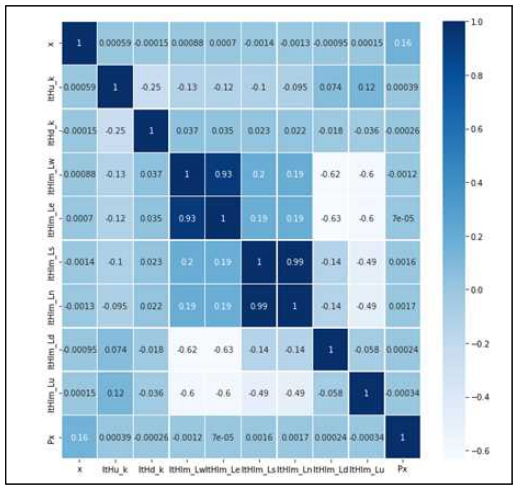

그림 5. tri_sor_dp_dp 변수의 상관계수
Fig. 5. Correlation Coefficient of Variable in tri_sor_dp_dp

각 모델의 학습은 tri_sor_dp_dp 함수에 사용되는 변수를 추출하여 9가지 Input feature를 통해 1가지 Output feature를 추정한다. 총 2000번 함수 호출 시 사용되는 데이터를 추출 후, 1999번 동안 함수 호출 된 데이터를 train dataset으로 2000번 째 함수호출 된 데이터를 test dataset으로 사용하였다. 각 함수 호출 1회당 10개의 feature가 각각 146880개의 데이터 값이 존재한다.

3.1 다층 퍼셉트론 회귀

다층 퍼셉트론(Multilayer Perceptron)은 입력층과 출력층에 한 개 이상의 완전 연결 은닉층(hidden layer)을 추가하고, 각 은닉층의 결과를 활성화 함수를 사용하여 예측을 진행하는 모델이다. 이때 최소 두 개 이상의 연산을 진행하는 층으로 구성되어 있어야 하며 비선형 함수를 추가하여 여러 개의 은닉층을 구성할 수도 있다. 다층 퍼셉트론 모델의 장점으로는 대량의 데이터에서 특징을 추출 가능하며, 복잡한 구성을 모델을 손쉽게 구현할 수 있다는 장점이 있지만, 학습 시간이 오래 걸리며, 모델의 튜닝이 어렵다는 단점이 존재한다[10].

다층 퍼센트론 회귀 모델에 학습을 진행 하기전 test dataset의 output feature인 변수 “Px”에 대하

여 IQR 기반 이상치 데이터를 제거 후 학습을 진행 하였다. 다층 퍼센트론 회귀 모델의 구성으로 Hidden Layer의 크기를 10개의 노드를 가진 2개의 Layer로 구성하였으며, 조기 종료를 설정하여 총 300번의 학습을 진행하였다. 가중치 최적화를 위한 함수로 adam 확률적 경사 기반 최적화를 사용하였으며, 활성화 함수로 Rectified Linear Unit을 사용하였다.

3.2 K-최근접 이웃 회귀

K-최근접 이웃(K-Nearest Neighbor, KNN)은 훈련 데이터셋에서 가장 가까운 이웃과 연관된 타겟을 국부 보간법을 사용하여 예측하는 모델이다. 객체는 k개의 최근접 이웃을 Distance Formula를 사용해 선택하며 선택된 k개의 이웃이 가진 평균값을 연산해 학습하는 greedy 학습을 진행한다. K-최근접 이웃 모델은 매우 간단하며, 학습 데이터의 분포를 고려하지 않아도 된다는 장점이 있다. 또한, 적은 데이터셋에서도 효과적으로 작동하며 여러 분야에 적용이 가능하다[11].

3.3 Light Gradient Boosting Machine

Light Gradient Boosting Machine (LightGBM)은 Tree 기반 머신러닝 알고리즘이며 Gradient Boosting 방식의 프레임워크 이다. 다른 트리 기반 알고리즘과 비교하였을 때 Light Gradient Boosting Machine은 트리를 수평으로 확장하는 방식이 아닌 left-wise tree growth 방식을 채택하여 수직으로 확장하는 방식으로 학습한다. 다른 level-wise 알고리즘과 비교하였을 때 낮은 오차를 보이는 특징이 있다. 학습에 소요되는 시간이 매우 빠르며 메모리를 적게 차지한다는 장점이 있지만 overfitting에 민감하다는 단점 또한 존재한다[12].

본 논문에서는 Light Gradient Boosting Machine 회귀 모델에 K Folds Cross Validation Method를 적용하여 검증하였다. 총 10개의 fold를 구성하여 교차 검증 하였다. 트리의 최대 잎의 개수는 31개, 최소 잎의 개수는 20개로 설정하였다. 트리의 최대 깊이는 제한 없으며 학습률은 0.1로 설정하였다.

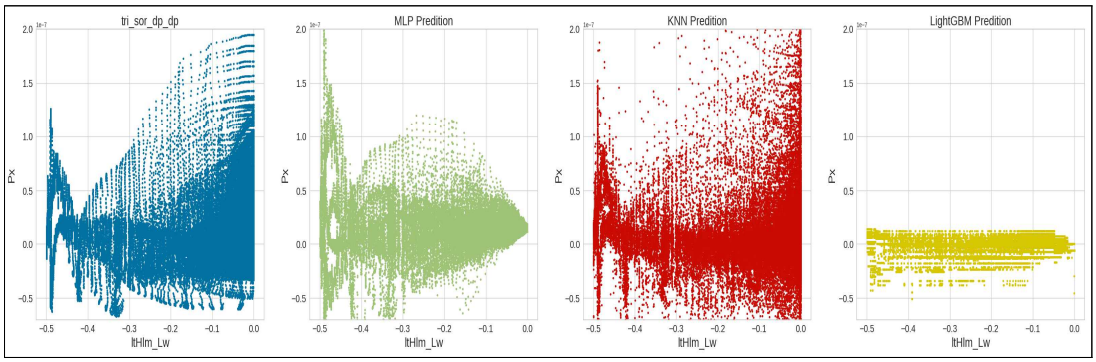

그림 6. Px 예측값
Fig. 6. Px prediction value

본 논문에서는 K-최근접 이웃 모델의 구성으로 이웃의 수를 3개로 설정하여 특정 값의 최근접 이웃 3개를 선택하여 평균을 구해 학습을 진행하였다. 각 지역의 모든 타겟에 균일한 가중치를 적용하도록 하였으며, 거리를 측정할때는 유클리디안 거리 방정식을 이용하여 계산하였다.

표 4. 실험에 수행된 시스템 스펙
Table 4. System specs used by test

Name	Hardware Specification
CPU	Intel® Core 13th GEN i9-13900F
RAM	128GB
SSD	2TB
GPU	GeForce RTX 4070TI

4. 실험 및 실험 결과

4.1 실험 환경

본 논문에서 머신러닝 수행에 사용된 시스템 스펙은 표 4와 같다. 머신러닝 학습 및 데이터 분석에 사용된 라이브러리로 python3.9.18, pandas, numpy, scikit-learn, matlibplot을 사용하였다. 다층 퍼셉트론, K-최근접 이웃, Light Gradient Boosting Machine으로 총 3가지 회귀 모델을 각각 구성하여 학습 후 성능 평가 및 비교를 진행하였다.

4.2 성능 평가

기존의 함수 호출에 적용하기 용이하도록 함수 호출 1회 데이터를 기준으로 예측 후 평가 및 비교를 진행하였다.

첫 번째로 그림 6과 같이 예측을 진행한 Output feature로 변수 “Px”를 tri_sor_dp_dp의 연산 값과 회귀 모델 세가지의 산포도 그래프를 각각 작성하여 비교하였다. 이때 x축의 기준은 tri_sor_dp_dp 함수의 Input feature 중 하나인 “ltHlm_Lw’ 변수를 선정하였다. 그래프를 살펴보았을 때 K-최근접 이웃 모델이 tri_sor_dp_dp 연산 값과 가장 유사한 모습을 보여 주며 다음으로 다층 퍼셉트론 회귀 모델이 가장 유사하며 Light Gradient Boosting Machine 회귀 모델은 그래프 상으로는 매우 다른 모습을 보여준다. 이러한 결과는 이웃 거리 평균을 이용하는 K-최근접 이웃 기법이기 때문에 해당 모델이 그래프 상으로 가장 유사한 모습을 나타내는 것으로 확인되었다.

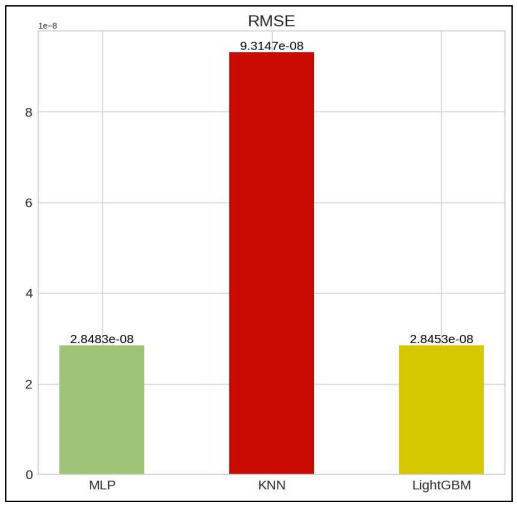

그림 7. 각 모델별 RMSE
Fig. 7. RMSE for each Model

두 번째로 그림 7과 같이 각 모델별 RMSE를 비교하였을 때 Light Gradient Boosting Machine 회귀 모델이 2.8453e-08으로 가장 낮게 측정 되었으며 그 다음으로 다층 퍼셉트론 회귀 모델이 2.8483e-08, K-최근접 이웃 회귀 모델이 9.3147e-08 순서로 측정되었다. 그래프 상에서 가장 유사하지 않았던 Light Gradient Boosting Machine 회귀 모델이 가장 우수한 RMSE로 측정 된 이유로는 측정되는 값이 매우 작은 단위이며 다른 예측과 비교하였을 때 0에 더 가깝기 때문에 이러한 결과가 측정되었다.

세 번째로 그림 8과 같이 각 모델별 MAE를 비교하였을 때 K-최근접 이웃 회귀 모델이 1.3637e-08로 가장 낮게 측정되었으며, 그다음으로 Light Gradient Boosting Machine 회귀 모델이 1.6605e-08, 다층 퍼셉트론 회귀 모델이 2.1321e-08로 가장 높은 오차로 측정되었다. 가장 우수한 성능을 보인 K-최근접 이웃 회귀 모델은 이웃한 값들의 거리 평균을 이용하는 특성 때문에 MAE 오차 지표에서 가장 우수한 성능을 보이는 것으로 측정되었다.

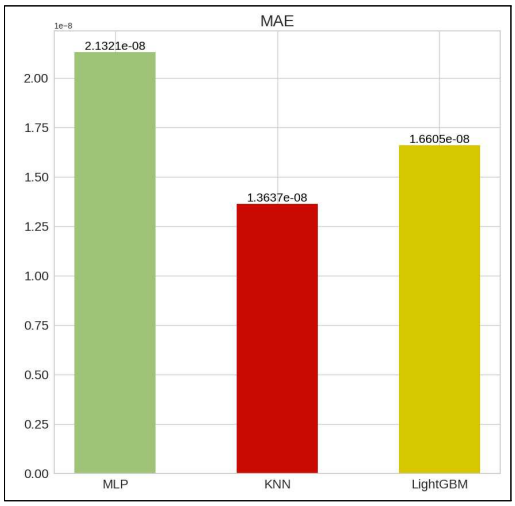

그림 8. 각 모델별 MAE
Fig. 8. MAE for each Model

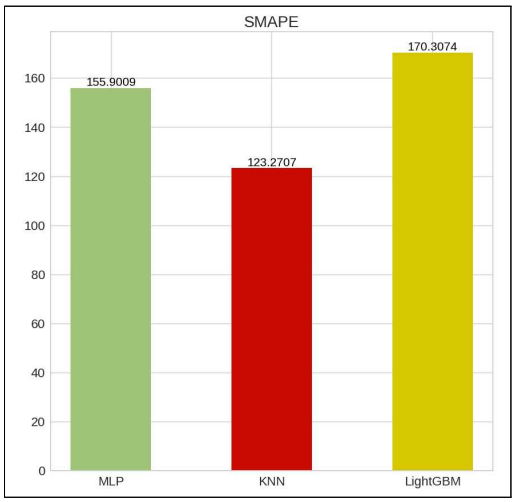

그림 9. 각 모델별 SMAPE
Fig. 9. SMAPE for each Model

네 번째로 그림 9와 같이 각 모델별 SMAPE를 비교하였다. SMAPE 본 논문에서 예측하는 값과 같이 데이터의 값이 작을 때 비율과 관련하기 때문에 결과 해석에 용이하며, 0~200% 사이의 확률값을 가지며 낮을수록 우수한 성능을 뜻한다. K-최근접 이웃 회귀 모델이 123.2707%로 가장 낮게 측정되었으며, 그다음으로 다층 퍼셉트론 회귀 모델이 155.9009%, Light Gradient Boosting Machine 회귀 모델이 170.3074%로 가장 높게 측정되었다. 가장 우수한 성능을 보인 K-최근접 이웃 회귀 모델은 산포도 상에서 실제 값과 가장 유사한 모습을 보여주었기 때문에 얼마나 유사하냐의 비율을 나타내는 SMAPE 오차 지표 특성상 가장 우수한 성능을 보이는 것으로 측정되었다.

5. 결론

본 논문에서는 중소규모 서버에서의 기상 연구를 위한 저해상도 결합모델인 Low-GloSea6의 Intel Vtune Profiler를 사용한 분석을 진행하였으며, 분석 결과로 전체 29877.390초의 CPU Time 중 1125.987초의 CPU Time을 소모하는 대기모델의 tri_sor.F90 모듈의 tri_sor_dp_dp 함수를 Hotspot으로 검출하였으며 모듈의 역할 및 데이터 구조를 분석하였다. 더욱 원활한 기상 연구를 위해 기존 수치적 연산을 진행하는 tri_sor_dp_dp 함수를 머신러닝 기법 중 하나인 비선형 회귀 모델 3가지를 구성하여 학습 후 평가 및 비교하였다. 평가를 진행한 모델 3가지는 다층 퍼셉트론, K-최근접 이웃, Light Gradient Boosting Machine을 구성하였으며 그 결과 K-최근접 이웃 회귀 모델이 MAE가 1.3637e-08, SMAPE가 123.2707%로 가장 우수하게 나타났으며 RMSE는 Light Gradient Boosting Machine 회귀 모델이 2.8453e-08로 가장 우수하게 측정되었다. 따라서 본 논문을 통하여 Low-GloSea6 수행 중 Hotspot으로 검출된 tri_sor_dp_dp 함수의 비선형 회귀 적용 가능성으로 가장 우수하게 나타난 모델은 K-최근접 이웃 회귀 모델로 해당 모델의 적용 가능성을 확인하였다.

REFERENCES

[1] D. W. Pierce, T. P. Barnett, B. D. Santer, P. J. Glecker, “Selecting global climate models for reginal climate change studies”, National Acad Sciences, vol. 106, no. 21, pp. 8441-8446, 2009
[2] KMA - Exchange of Memorandum of

Understanding between the National Center for Atmospheric Research and Korea Meteorological Administration, [online] https://www.kma.go.kr/kma/news/press.jsp

[3] H. Kim, J. Lee, Y. Hyun, S. Hwang, "The KMA Global Seasonal Forecasting System (GloSea6) - Part 1: Operational System and Improvements", Korean Meteorological Society, vol. 31, no. 3, pp. 341-359, 2021.

[4] Y. Hyun, J. Lee, B. Shin, Y. Choi, J. Kim, S. Lee, H. Ji, K. Boo, S. Lim, H. Kim, Y. Ryu, Y. Park, H. Park, S. Choo, S. Hyun, S. Hwang, "The KMA Global Seasonal forecasting system (GloSea6) - Part 2: Climatological Mean Bias Characteristics", Korean Meteorological Society, vol. 33, no. 2, pp. 87-101, 2022.

[5] H. Park, Y. Cho, D. Shin, E. Yun, S. Chung, "A Study of the Application of Machine Learning Methods in the Low-GloSea6 Weather Prediction Solution", The Journal of Korea Institute of Information, Electronics, and Communication Technology, vol. 16, no. 5, pp. 307-314, 2023

[6] L. Chen, F. Du, Y. Hu, Z. Wang, F. Wang, "SwinRDM: Integrate SwinRNN with Diffusion Model towards High-Resolution and High-Quality Weather Forecasting", The Thirty-Seventh AAAI Conference on Artificial Inteligence, vol. 37, no. 36, pp. 322-330, 2023

[7] S. Yukimoto, Y. Adachi, M. Hosaka, T.Sakami H. Yoshimura, M. Hirabara, T. Y. Tanaka, E. Shindo, H. Tsujino, M. Deushi, R. Mizuta, S. Yabu, A. Obata, H. Nakano, T. Koshiro, T. Ose, A. Kitoh, "A New Global Climate Model of the Metorological Research Institute: MRI-CGCM3-Model Description and Basic Performance-", Meteorological Society of Japan, vol. 90, pp. 23-64, 2012.

[8] VTune-profiler [online] https://www.intel.com/content/www/us/en/developer/tools/oneapi/vtune-profiler.html#gs.5xzf5j

[9] VTune-profiler [online] https://www.intel.com/content/www/us/en/docs/vtune-profiler/user-guide/2024-0/tuning-methodology.html

[10] Karlik, B. and Olgac, A.V, "Performance Analysis of Various Activation Functions in Generalized MLP Architectures of Neural Networks", International Journal of Artificial Intelligence And Expert Systems (IJAE), vol. 1, no. 4, pp. 111-122, 2011.

[11] O. Kramer, "K-nearest neighbors", Dimensionality Reduction with Unsupervised Nearest Neighbors, vol. 51, pp. 13-23, 2013.

[12] J. Fan, X. Ma, L. Wu, F. Zhang, X. Yu, W. Zeng, "Light Gradient Boosting Machine: An efficient soft computing model for estimating daily reference evapotranspiration with local and external meteorological data", Agricultural water management, vol. 225, no. 20, 2019.

저자약력

박 혜 성 (Hye-Sung Park) **[학생회원]**

- 2020년 3월 ~ 현재: 창원대학교 학부 과정

〈관심분야〉 AI, 머신러닝, 데이터 분석

조 예 린 (Ye-Rin Cho) **[학생회원]**

- 2020년 3월 ~ 현재: 창원대학교 학부 과정

〈관심분야〉 IoT, 실시간 분산 멀티미디어

신 대 영 (Dae-Yeon Shin) **[학생회원]**

- 2020년 3월 ~ 현재: 창원대학교 학부 과정

〈관심분야〉 IoT, 실시간 분산 멀티미디어

윤 은 옥 (Eun-Ok Yun) **[학생회원]**

- 2020년 3월 ~ 현재: 창원대학교 학부 ,과정

〈관심분야〉 유무선 네트워크, IoT

정 성 욱 (Sung-Wook Chung) **[종신회원]**

- 2010년 8월: CISE dept. Univ. of Florida, USA, (Ph.D)
- 2010년 10월 ~ 2012년 2월: KT 종합기술원 중앙연구소 선임연구원
- 2012년 3월 ~ 현재: 창원대학교 컴퓨터공학과 정교수

〈관심분야〉 IoT, 스마트모빌리티, HPC, 실시간 분산 멀티미디어 시스템