The Scientific Temper (2024) Vol. 15 (spl): 146-149

E-ISSN: 2231-6396, ISSN: 0976-8653

Doi: 10.58414/SCIENTIFICTEMPER.2024.15.spl.18

https://scientifictemper.com/

RESEARCH ARTICLE

Enhanced regression method for weather forecasting

T. Malathi[1*], T. Dheepak[2]

Abstract

Weather prediction is gaining popularity very rapidly in the current era of artificial intelligence and Technologies. It is essential to predict the temperature of the weather for some time. Traditionally, weather predictions are performed with the help of large complex models of physics, which utilize different atmospheric conditions over a long period of time. These conditions are often unstable because of perturbations of the weather system, causing the models to provide inaccurate forecasts. The models are generally run on hundreds of nodes in a large high-performance computing (HPC) environment, which consumes a large amount of energy. In this paper, LightGBM Regression parameters are tuned by using an optimization technique. Differential evolution (DE) is used to optimize the LightGBM regressor for estimating and forecasting the weather in the fore coming days.

Keywords: Weather forecasting, Light gradient boosting machine, Regression, Differential evolution.

Introduction

Predictions for the future using the correct algorithm are viral nowadays. This prediction is applicable for the weather prediction as well. It can use machine learning to know whether it will rain tomorrow or what the temperature will be tomorrow. Machine learning algorithms can correctly forecast weather features like humidity, temperature, outlook, and airflow speed and direction. This sector is immensely dependent on previous data and artificial intelligence. Predicting future weather also helps us to make decisions in agriculture, sports and many aspects of our lives. Weather conditions around the world change rapidly and continuously. Correct forecasts are essential in today's daily life. From agriculture to industry, from traveling to daily commuting, the people are dependent on weather forecasts heavily. As the entire world is suffering from the continuous climate change and its side effects, it is very important to predict the weather without any error to ensure easy and seamless mobility, as well as safe day to day operations, Shiu, Y. S., & Chuang, Y. C. (2019), Maleki, M., Barkhordar, Z., Khodadadi, Z., & Wraith, D. (2019), Bun, M. J., & Harrison, T. D. (2019), Gelman, A., Goodrich, B., Gabry, J., & Vehtari, A. (2019), Ventura, M., Saulo, H., Leiva, V., & Monsueto, S. (2019), Yousof, H. M., Altun, E., Rasekhi, M., Alizadeh, M., Hamedani, G. G., & Ali, M. M. (2019), Sabottke, C. F., Breaux, M. A., & Spieler, B. M. (2020).

The current weather prediction models heavily depend on complex physical models and need to be run on large computer systems involving hundreds of HPC nodes. The computational power of these large systems is required to solve the models that describe the atmosphere. Despite using these costly and complex devices, there are often inaccurate forecasts because of incorrect initial measurements of the conditions or an incomplete understanding of atmospheric processes. Moreover, it generally takes a long time to solve complex models like these. Through this research paper, an optimization-based regressor is proposed to enhance weather forecasting in the upcoming days, Chen, X., Huang, J., & Yi, M. (2020), Jabeur Telmoudi, A., Soltani, M., Chaouech, L., & Chaari, A. (2020), Kipourou, D. K., Charvat, H., Rachet, B., & Belot, A. (2019).

Differential Evolution Optimization

Differential evolution is found to be a well-known evolution-based optimization technique. It has elegantly combined

[1]Assistant Professor, Department of Computer Applications, Shrimati Indira Gandhi College (Affiliated to Bharathidasan University, Tiruchirappalli) Tiruchirappalli -620002, Tamil Nadu, India.

[2]Assistant professor of Computer science CDOE, Bharathidasan University, Tiruchirappalli-620 024, Tamil Nadu, India.

***Corresponding Author:** T. Malathi, Assistant Professor, Department of Computer Applications, Shrimati Indira Gandhi College (Affiliated to Bharathidasan University, Tiruchirappalli) Tiruchirappalli -620002, Tamilnadu, India, E-Mail: mailto:malathits@gmail.com

How to cite this article: Malathi, T., Dheepak, T. (2024). Enhanced regression method for weather forecasting. The Scientific Temper, **15**(spl):146-149.

Doi: 10.58414/SCIENTIFICTEMPER.2024.15.spl.18

Source of support: Nil

Conflict of interest: None.

Received: 17/09/2024 **Accepted:** 28/09/2024 **Published**: 16/10/2024

the mutation and crossover operators of GA. The major difference between GA and differential evolution is that it requires fewer control parameters. Therefore, it has better convergence speed as compared to GA. Differential evolution can improve the parameters selection process. Generally, differential evolution consists of four main steps (i.e. initializing population, mutation and recombination, selection, and stopping criteria) to find optimal solutions and it is represented in Figure 1, Mohamed, A. W., & Mohamed, A. K. (2019), Tyralis, H., & Papacharalampous, G. (2017), Karasu, S., & Altan, A. (2019, November), Malathi, T., & Manimekalai, M. (2020).

Step 1: Initializing population

A given set of random solutions is developed in this step, considering that each parameter lies within the range of lower and upper bound. The normal distribution is used to develop these random values.

Step 2: Mutation and recombination

The primary benefit of differential evolution is the way it generates solutions throughout its life cycle. The difference between any two solutions of differential evolution is added with third solution to form a new solution. Therefore, it can elegantly overwrite the mutation and crossover operators of GA. This proves that differential evolution can converge at a faster speed as compared to GA.

Step 3: Selection

The fitness value of solutions is then evaluated using a given objective function. In case if the fitness of the new solution is more than that of the best-known fitness so far, then the best-known solution will be changed to a new solution and continue otherwise for further generations.

Step 4: Stopping Criteria

It is not possible to achieve maximum fitness function. Therefore, one can use either number of iterations, number of function evaluations, or acceptance error (AE), to finish the differential evolution procedure.

Light Gradient Boosting Machine (Lightgbm) Regression Technique

The LightGBM model is an ensemble learning model based on gradient facilitated decision tree (GBDT). The key point of the model is to accumulate all three models as the output result. The LightGBM model is optimized in the original GBDT algorithm, which solves the problem of the difficulty of training large amounts of data in GBDT. The traditional GBDT-based algorithm such as Xgboost, uses a pre-sorting method for prediction. First, all values on the data set are pre-sorted according to the features, and the best segmentation point on the feature is found by traversing the entire data set. This greatly increases the time complexity and memory usage. The LightGBM model uses the histogram algorithm, the

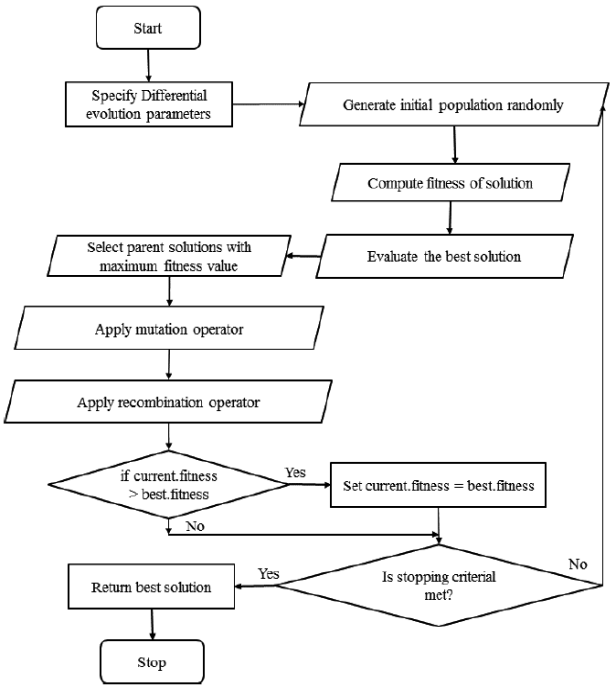

Figure 1: Flowchart of differential evolution optimization algorithm

continuous eigenvalues are discretized into N integers and a histogram with a unit of N is constructed. When traversing the data, the discrete eigenvalues are used as an index in the histogram. Calculate the statistics in the middle, and finally, find the optimal split point according to the discrete value of the histogram. The LightGBM model also uses the idea of mutually exclusive feature bundling (EFB), which combines some mutually exclusive features to reduce the dimension of features. Finally, the LightGBM model adopts the leaf-wise leaf growth strategy with depth limitation. Unlike the GBDT algorithm, it does not need to search and split each layer of leaves. In this way, the operating efficiency of the computer is greatly improved, Roozbeh, M., Hesamian, G., & Akbari, M. G. (2020), Almeshal, A. M., Almazrouee, A. I., Alenizi, M. R., & Alhajeri, S. N. (2020), Alves, R. S., de Resende, M. D. V., Azevedo, C. F., Silva, F. F. E., Rocha, J. R. D. A. S. D. C., Nunes, A. C. P., ... & dos Santos, G. A. (2020).

Optimization Based Regressor Model

In this proposed enhanced regression technique, differential evolution to optimize the LightGBM tree in such a way that it maximizes the specificity and sensitivity ratio of weather data, Marcoulides, K. M., & Raykov, T. (2019), Deng, W., Shang, S., Cai, X., Zhao, H., Song, Y., & Xu, J. (2021).

Step 1: Initialization

In this step, DE procedure is initiated by developing its initial population randomly using Normal Distribution with Mean =0, and variance =1. Each random solution has a size 17. The lower bound and upper bound values are between 0 and 1, respectively.

Step 2: Fitness Function

The DE-LGBM fitness function is defined by minimizing the following symmetric mean absolute percentage error (SMAPE).

$$\left\{\min f_1(z) = \frac{\sum_{t=1}^{n}|F_t - A_t|}{\sum_{t=1}^{n}(F_t + A_t)}\right.$$

F_t is the forecasted value for given time series, and A_t is the actual values for given time series.

Step 3: Selection Operator

The SMAPE is calculated. Variable V1 stores the best low smape ratio of solution so far. Every random solution corresponding to the best smape ratio store into a matrix best solution.

Step 4: Mutation and recombination operator

Differential evolution develops trial solutions in an efficient way using mutation and recombination operators. A weighted difference between two solutions is merged into a third solution to generate a new solution. The fitness of the developed solution is also calculated to compare it with the best solution so far. In case the generated solution has a significant fitness value, then best solution will be replaced with this solution. Continue otherwise.

Step 5: Termination criteria

This step will repeat all its steps till the best fitness value of each solution is less than to threshold is defined. Once any best fitness value greater or equal to the threshold is found, then DE will automatically return optimized parameters.

Table 1: Performance metrics for optimization-based regressor model

Error Name	*Equation*
Symmetric mean absolute percentage error (SMAPE)	$\frac{\sum_{t=1}^{n}\|F_t - A_t\|}{\sum_{t=1}^{n}(F_t + A_t)}$
Mean absolute percentage error (MAPE)	$\frac{1}{n}\sum_{t=1}^{n}\left\|\frac{A_t - F_t}{A_t}\right\|$
Root mean squared error (RMSE)	$\sqrt{\frac{\sum_{t=1}^{n}(F_t - A_t)^2}{n}}$

Table 2: Performance analysis of the proposed OR model and LightGBM model using the original dataset and proposed AC-IG processed dataset

Performance metrics	*Regression models*	
	Proposed OR model	*LightGBM model*
	Original Dataset	
SMAPE	42.43	45.76
MAPE	41.69	46.92
RMSE	38.66	41.72
	Proposed AC-IG based FS processed dataset	
SMAPE	18.86	22.26
MAPE	31.72	41.64
RMSE	24.58	37.28

Result And Discussion

Performance Metrics

Table 1 gives the performance metrics used in this contribution to evaluate the performance of the proposed optimization-based regressor (OR) model.

The performance of the Proposed OR model is compared with the existing LightGBM regressor using original dataset and the proposed AC-IG-based feature selection method [] processed dataset.

Table 2 gives the performance analysis of the proposed OR model using the original dataset and the proposed AC-IG method processed dataset. Figure 2 depicts the graphical representation of the symmetric mean absolute percentage error (SMAPE) with the proposed OR model and LightGBM model using the original dataset and proposed AC-IG processed dataset. From Table 2 and Figure 2, it is clear that the proposed OR model with proposed AC-IG-based feature Selection processed dataset gives less SMAPE value when it is compared with the existing model.

Figure 3 depicts the graphical representation of the Mean Absolute Percentage Error (MAPE) with the proposed OR model and LightGBM model using the original dataset and proposed AC-IG processed dataset. From Table 2 and Figure 3, it is clear that the proposed OR model with the proposed AC-IG processed dataset gives less MAPE when it is compared with the LightGBM model.

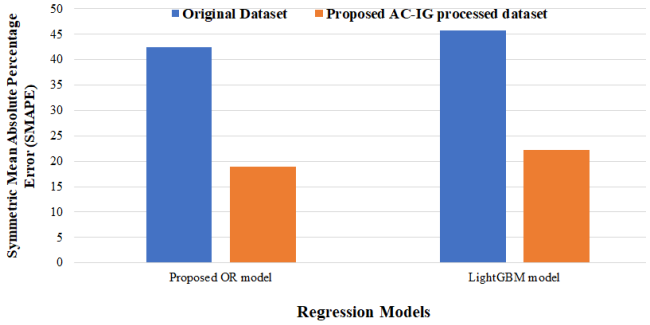

Figure 2: Graphical representation of the SMAPE with the proposed OR model and LightGBM model using original dataset and proposed AC-IG processed dataset

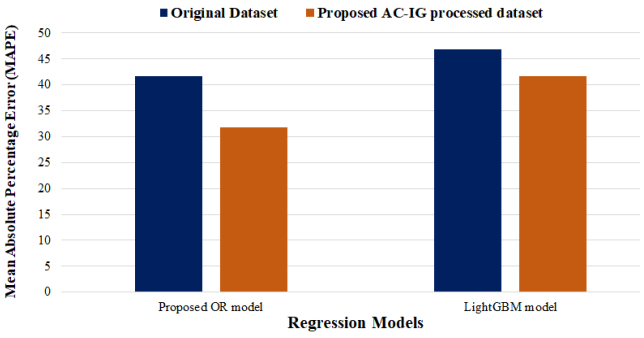

Figure 3: Graphical representation of the MAPE with proposed OR model and LightGBM model using the original dataset and proposed AC-IG processed dataset

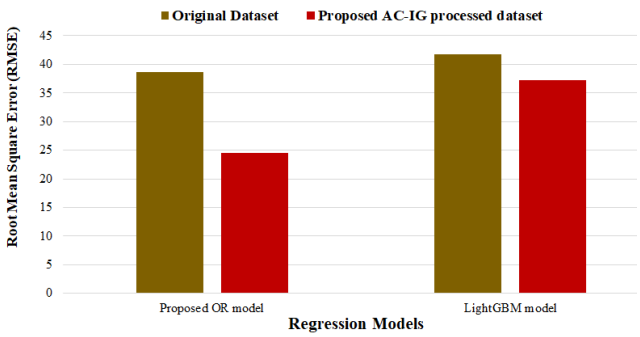

Figure 4: Graphical representation of the RMSE with proposed OR model and LightGBM model using the original dataset and proposed AC-IG processed dataset

Figure 4 depicts the graphical representation of the Root Mean Squared Error (RMSE) with the proposed OR model and LightGBM model using the original dataset and proposed AC-IG processed dataset. From Table 2 and Figure 4, it is clear that the proposed OR model with the proposed AC-IG processed dataset gives less RMSE when it is compared with LightGBM model.

Conclusion

In this chapter, an optimization based regression model is proposed with LightGBM and differential evolution optimization. Through this proposed OR model, the forecasting of the weather can be prediction for future days. The performance of the proposed OR model is compared with existing regressors like LightGBM for the original dataset and the proposed AC-IG-based feature selection processed dataset.

References

Almeshal, A. M., Almazrouee, A. I., Alenizi, M. R., & Alhajeri, S. N. (2020). Forecasting the spread of COVID-19 in Kuwait using compartmental and logistic regression models. Applied Sciences, 10(10), 3402.

Alves, R. S., de Resende, M. D. V., Azevedo, C. F., Silva, F. F. E., Rocha, J. R. D. A. S. D. C., Nunes, A. C. P., ... & dos Santos, G. A. (2020). Optimization of Eucalyptus breeding through random regression models allowing for reaction norms in response to environmental gradients. Tree Genetics & Genomes, 16, 1-8.

Bun, M. J., & Harrison, T. D. (2019). OLS and IV estimation of regression models including endogenous interaction terms. Econometric Reviews, 38(7), 814-827.

Chen, X., Huang, J., & Yi, M. (2020). Cost estimation for general aviation aircrafts using regression models and variable importance in projection analysis. Journal of cleaner production, 256, 120648.

Deng, W., Shang, S., Cai, X., Zhao, H., Song, Y., & Xu, J. (2021). An improved differential evolution algorithm and its application in optimization problem. Soft Computing, 25, 5277-5298.

Gelman, A., Goodrich, B., Gabry, J., & Vehtari, A. (2019). R-squared for Bayesian regression models. The American Statistician.

Jabeur Telmoudi, A., Soltani, M., Chaouech, L., & Chaari, A. (2020). Parameter estimation of nonlinear systems using a robust possibilistic c-regression model algorithm. Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering, 234(1), 134-143.

Karasu, S., & Altan, A. (2019, November). Recognition model for solar radiation time series based on random forest with feature selection approach. In 2019 11th international conference on electrical and electronics engineering (ELECO) (pp. 8-11). IEEE.

Kipourou, D. K., Charvat, H., Rachet, B., & Belot, A. (2019). Estimation of the adjusted cause-specific cumulative probability using flexible regression models for the cause-specific hazards. Statistics in medicine, 38(20), 3896-3910.

Malathi, T., & Manimekalai, M. (2020). Feature selection techniques for weather forecasting models using machine learning techniques. Journal of Electrical Engineering and Technology, 11(4), 443-455.

Maleki, M., Barkhordar, Z., Khodadadi, Z., & Wraith, D. (2019). A robust class of homoscedastic nonlinear regression models. Journal of Statistical Computation and Simulation, 89(14), 2765-2781.

Marcoulides, K. M., & Raykov, T. (2019). Evaluation of variance inflation factors in regression models using latent variable modeling methods. Educational and psychological measurement, 79(5), 874-882.

Mohamed, A. W., & Mohamed, A. K. (2019). Adaptive guided differential evolution algorithm with novel mutation for numerical optimization. International Journal of Machine Learning and Cybernetics, 10, 253-277.

Roozbeh, M., Hesamian, G., & Akbari, M. G. (2020). Ridge estimation in semi-parametric regression models under the stochastic restriction and correlated elliptically contoured errors. Journal of Computational and Applied Mathematics, 378, 112940.

Sabottke, C. F., Breaux, M. A., & Spieler, B. M. (2020). Estimation of age in unidentified patients via chest radiography using convolutional neural network regression. Emergency radiology, 27, 463-468.

Shiu, Y. S., & Chuang, Y. C. (2019). Yield estimation of paddy rice based on satellite imagery: Comparison of global and local regression models. Remote Sensing, 11(2), 111.

Tyralis, H., & Papacharalampous, G. (2017). Variable selection in time series forecasting using random forests. Algorithms, 10(4), 114.

Ventura, M., Saulo, H., Leiva, V., & Monsueto, S. (2019). Log-symmetric regression models: information criteria and application to movie business and industry data with economic implications. Applied Stochastic Models in Business and Industry, 35(4), 963-977.

Yousof, H. M., Altun, E., Rasekhi, M., Alizadeh, M., Hamedani, G. G., & Ali, M. M. (2019). A new lifetime model with regression models, characterizations and applications. Communications in Statistics-Simulation and Computation, 48(1), 264-286.