International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 02 | February - 2023 Impact Factor: 7.185 ISSN: 2582-3930

CLASSIFICATION FOR BIG DATA DRIVEN MARINE WEATHER FORECASTING USING MACHINE LEARNING TECHNIQUES

Deepa Anbarasi J[1], Dr. V. Radha[2]

[1]Research Scholar, Department of computer science, Avinashilingam institute of home science and higher education for women, Coimbatore.

[2]Proffoser, Department of computer science, Avinashilingam institute of home science and higher education for women, Coimbatore.

ABSTRACT

Marine weather forecasting has raised sizeable awareness in numerous ocean related domains. With immeasurable portion of data to handle with, big data problem solving unfolds gateways for abundance of predictions. Machine Learning (ML) is an essential algorithm for big data prediction. However with the increasing size and nature of data i.e., Big Data, Marine Weather Forecasting with Big Data with minimum time, error and maximum accuracy is of major concern to be addressed. In this work, a method called, Perceptred-based Feature and Kriging Gradient Boost Classification (PF-KGBC) is introduced with big data with the objective of improving the prediction performance marine weather with high accuracy and less time consumption. The PF-KGBC method is split into two parts. They are feature selection using perceptron classifier model and classification using Kriging Ensemble extreme Gradient Boost for marine weather forecasting. With the assistance of supervised learning algorithm based on perceptron classifier that involves a functional input represented by vector of numbers belongs to particular class. Based on linear predictor function set of weights are integrated with feature vector to make essential marine weather feature vector. After feature selection process, Kriging Ensemble extreme Gradient Boost Classification is performed with the purpose of forecasting marine weather data. Here, the weak learner is combined to form strong classifier. Kriging regression estimates the dependent data variation when any of factors or values in independent data gets changed. With this accurate marine weather prediction with high accuracy and lesser time consumption is ensured. Meanwhile, we analyzed the results of the proposed PF-KGBC method compared with other conventional methods, and the running performance on the Java platform. The results show that the proposed method achieved satisfactory prediction results and improvements were observed in terms of accuracy, time and error rate considerably.

Keywords: Machine Learning, Perceptron, Feature Selection, Kriging, Gradient Boost, Classification, Ensemble

1. Introduction

Big data refers to the analysis and extracting information from large database, usually a combination of structured, semi-structured and unstructured data. By applying ML algorithms several domains are focused, like, weather forecasting, speech recognition, traffic prediction, online fraud detection and so on.

Among them, weather forecasting predicts the state of atmosphere for specific location on the basis of distinct weather factors. Hence, accurate weather forecasting remains to be the challenging issue for meteorologists and researchers. On the other hand, with marine weather forecasting, mariners and meteorological organizations forecast the future weather conditions over Earth oceans. Several researchers carried out their research on marine weather forecasting with big data.

A semi-supervised prediction method called, Spark-based fuzzy partitioning LSTM networks (SFPLN) was introduced in [1] with the assistance of improved unsupervised clustering algorithm. With the aid of this algorithm assisted in identifying fuzzy partition function. Followed by which a neural network model was employed to

construct information prediction function. With this method, the time consumption and error involved for forecasting marine information was said to be considerably reduced. Despite improvement observed in terms of both time consumption and error, the prediction accuracy involved in the marine forecasting was not focused.

Multilayer Convolutional Long and Short Term Memory (M-convLSTM) was designed in [2] to forecast the 3-D ocean temperature. The designed model included the convolutional neural networks (CNNs), long-and short-term memory (LSTM) and multiple layer stacking with horizontal and vertical temperature variations from sea surface to subsurface about 2000 m below. With this the prediction accuracy of ocean temperature prediction was said to be improved significantly. Despite improved found in the prediction accuracy, the prediction time involved in ocean temperature was not concentrated.

Deep Learning Long Short Term Memory (DL-LSTM) Neural Network that performed location-specific Sea Surface Temperature (SST) forecasts was carried out in [3] by integrating deep learning neural networks with numerical estimators at five locations for three distinct time horizons (daily, weekly and monthly). Neural Networks enhanced their result performance by deep learning long short-term memory (LSTM) neural network. However, the accuracy level was not improved.

1.1 Problem Definition

Recent developments in marine weather forecasting have led to the growth of different new machine learning based techniques. In the machine learning based techniques, a classification algorithm is trained by employing certain features that can predict marine weather at an early stage in a precise manner. These features are extracted from the dataset. The existing methods predict weather condition using fuzzy layer, multiple layer stacking and neural networks. However, with the nature and size of data increasing due to the introduction of big data, they are complicated, slow in nature, hence compromising accuracy and time. Therefore, there is requirement of marine weather forecasting at a timely and accurately for Big Data solution.

1.2 Proposed Solution

To solve above said problem this paper proposes a novel Big Data driven marine weather forecasting method that avoids the aforementioned shortcomings and tries to incorporate a classification algorithm using machine learning technique. At the heart of the method are the social aspects of oceanographic and surface meteorological readings for marine weather forecasting at an early stage and avoid the damaged caused by it. Such sociological marine weather forecasting aspects have been revealed in several studies [1]–[2], and the implementation of this idea was based on deep learning techniques.

In particular, this paper develops two distinct processes, Linear Perceptron-based Feature Selection and Kriging Ensemble extreme Gradient Boost Classification. First, Linear Perceptron-based Feature Selection is used to select the most pertinent and relevant features for classification. Second, Kriging Ensemble extreme Gradient Boost Classification is used to perform classification based on the ensemble resultant values of kriging regression, therefore boosting the overall performance ensuring time and accurate marine weather forecasting.

1.3 Contributions

The followings are the major contributions of our paper:

- To design a proposed method called Perceptred-based Feature and Kriging Gradient Boost Classification (PF-KGBC) with big data for marine weather forecasting.
- Proposed method selects the most pertinent and relevant features from the E1 Nino dataset by applying Linear Perceptron-based Feature Selection model. Therefore, it can improve the prediction accuracy by also ensuring minimum prediction time.
- Proposed method classifies the weather data accurately by employing Kriging Ensembled eXtreme Gradient Boost Classification algorithm that with the aid of Kriging regression estimates Multivariate Regressions for modeling ensemble data and then by applying transformation for Gradient Boosting also reduces the error rate.

- We have also conducted a experimental analysis for marine weather forecasting using parameters like, prediction accuracy, prediction time and error rate.

1.4 Outlines

The rest of this paper is organized as follows. Section 2 describes the literature work on weather forecasting, Big Data, machine learning techniques, deep learning techniques and distinct classification methods relevant to this paper. In Section 3, the Perceptred-based Feature and Kriging Gradient Boost Classification (PF-KGBC) method is described. Section 4 presents the simulation environment and results. Finally, Section 5 contains the conclusion.

2. Related works

In this section we present an overview of marine weather forecasting solutions proposed in the literature. One of the most predominant meteorological factors affecting in many facets of our lives is rainfall. With the facets varying from destruction to infrastructure in the occurrence of a flood to interferences in the network, therefore the socio-economic effects of rainfall are significant. More compellingly, recent studies have underscored that weather conditions can likely increase air pollution.

Modeled surface currents were employed in [4] to measure the significance of interaction between wave and current therefore improving the accuracy. An elaborate comparative analysis [5] utilizing simplified rainfall estimation models on the basis of traditional Machine Learning and Deep Learning techniques for rainfall prediction. A study to analyze performance of several Numerical Weather Prediction (NWP) methods for forecasting Tropical Cyclone (TC) was proposed in [6], therefore causing an improvement in accuracy.

With a plethora of satellites and remote-sensing devices keeping an eye on weather systems globally, over the past few years, meteorologists now have possess data accessible to them than ever before. However, large data need not as a matter of course interpret into enhanced prediction results. Owing to the increase in size and complexity involving weather patterns, prediction yet remains to be a more cumbersome process.

A multimodal fusion method to construct weather visibility prediction system was proposed in [7]. Here, also an advanced numerical prediction for emission detection was utilized to construct multimodal fusion visibility prediction system. Moreover, novel regression algorithm, like XGBoost, and LightGBM were also employed for numerical prediction. However, only single parameter was employed namely Sea Surface Temperature (SST).

A deep learning method using Multi-Layer Perceptron (MLP) with Multi-Variant Convolutional (MVC) for predicting four significance metrics, like, temperature, pressure, salinity and density for efficient prediction was designed in [8]. An example system that has been employed for utility specific applications while addressing the Big Data issues for variable generation power forecasting was proposed in [9].

Over the past few years several numerical methods are currently utilized for forecasting weather pursuits to minimize the risks of numerous disasters related to hydro-meteorological data. A Weather Research and Forecasting (WRF) model capable of predicting the potentialities over the Italian national territory in 2018 using fuzzy was proposed in [10].

A mechanism to improve coastal fog prediction was designed in [11]. However, the method only involved long term prediction. To address prediction for farmers involving short term prediction, a lightweight and novel weather forecasting system, using machine learning techniques for weather forecasting was proposed in [12], therefore ensuring forecasting accuracy even for short term prediction.

Weather forecasting is an indispensable element of numerous hydrological studies. A comparison between three different machine learning methods, k-nearest neighbours (KNN), Soil and Water Assessment Tools (SWAT), and Representative Concentration Pathway (RCP) was elaborated in [13]. A survey on weather forecasting using deterministic parameters on the basis of intelligent predictors was proposed in [14]. Convolutional LSTM was applied in [15] for forecasting weather events.

In [16], a novel lightweight data-driven weather forecasting method by employing temporal modeling techniques both for Long Short Term Memory (LSTM) and Temporal Convolutional Networks (TCN) was proposed. Moreover, Arbitrage of Forecasting Expert (AFE) was also designed as dynamic ensemble model, followed by which time-series data were utilized to obtain information pertaining to weather. Also, distinct numbers of layers were employed over a given period of time for forecasting weather accurately.

The article in [17] discussed the characteristic of modeling weather conditions for marine weather forecasting by utilizing copula-based method. However, the prevailing prediction method of visibility is specifically based on prediction using numerical values moreover same as performing weather prediction.

A machine learning technique based on data acquired for a period of six years with respect to meteorological and pollution data for predicting concentrations of PM2.5 from wind speed, direction and distinct precipitation levels were proposed in [18]. Data assimilation was performed in [19] to analyze the impact of parameter in weather forecasting. However, improvement was not obtained in terms of error rate. To concentrate on this issue, time series data were utilized and applied to the K-Means clustering algorithm [20] to obtain the forecasting data. Also, a learning-based procedure was employed to evaluate the parameters for forecasting. With this forecasting was made with minimum error rate.

Motivated by the above issues, in this work, marine weather forecasting with big data is performed by employing Perceptred-based Feature and Kriging Gradient Boost Classification (PF-KGBC) method to perform accurate forecasting with minimum time and error. The elaborate description of the PF-KGBC method is described in the following sections.

3. Methodology

The oceans are an essential portion of the Earth's environment, dispensing food, and life. In spite of that, heterogeneity of human enactments is placing our oceans at risk. All over the place, source of pollution on land occurs from 80% of and coastal areas are hence specifically susceptible to contaminants. Nevertheless, in the prediction of weather conditions, numerous issues are experienced, like arbitrary and unstable features of wind and waves. To predict environmental and marine weather conditions in an accurate manner, several techniques have been proposed by researchers.

In our work, Perceptred-based Feature and Kriging Gradient Boost Classification (PF-KGBC) method is introduced for big data to address the issues related with accuracy and time consumption. Figure 1 shows the block diagram of PF-KGBC method.

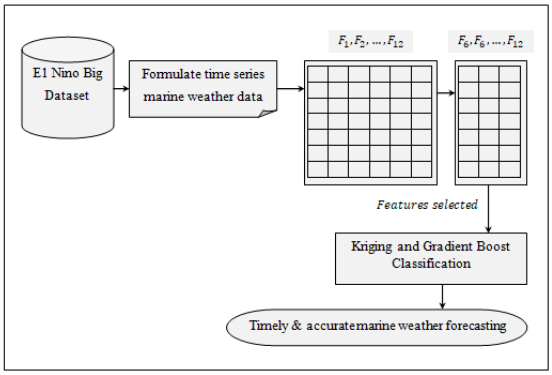

Figure 1 Block diagram of Perceptred-based Feature and Kriging Gradient Boost Classification

As shown in the above figure, the PF-KGBC method is split into two sections. The first section performs the feature selection using Linear Perceptron-based Feature Selection model. The second section performs classification for marine weather forecasting using Kriging Ensemble eXtreme Gradient Boost Classification-based Marine Weather Forecasting model. The elaborate description of the PF-KGBC method is given in the following sections.

3.1 Linear Perceptron-based Feature Selection

Marine weather forecasting is important to mariners and seamen, however prompt weather prediction can assist in preventing accidents results in cargo losses and even fatalities to some extent. In this section, Linear Perceptron-based Feature Selection (LPFS) is proposed for marine weather forecast. The main objective of this experiment is to predict marine weather with meteorological time series data as input

variable using LPFS model. Prior to prediction, the LPFS feature selection is utilized for identification of important features for marine weather prediction and reduces the time of the model. Figure 2 shows the structure of Linear Perceptron-based Feature Selection model.

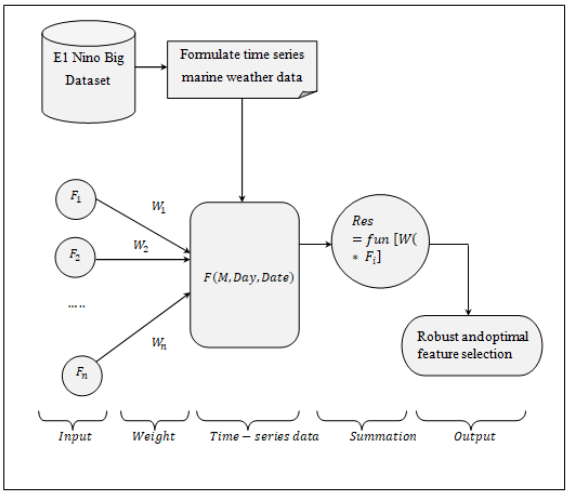

Figure 2 Structure of Linear Perceptron-based Feature Selection model

As shown in the above figure, perceptron classifier is a supervised learning algorithm of binary classifiers employed for performing feature selection process. The Linear Perceptron-based Feature Selection modeling task has been underscored prior to the defining of a network structure that includes time-series marine weather data sequence '$F = F_1, F_2, \dots, F_n$' and its objective remains in predicting the weather forecasting outputs '$Y = Y_1, Y_2, \dots, Y_n$' involving E1 Nino big data dataset 'DS' at each time.

Table 1 Weather feature details

S. No	Parameters
1	Observation
2	Year
3	Month
4	Day
5	Date
6	Latitude
7	Longitude
8	Zonal winds
9	Meridional winds
10	Humidity
11	Air temperature
12	Sea surface temperature

As given in the above table 1, there are 12 distinct weather parameters or features in E1 Nino big data dataset at a given time 't', '$F_t = F_1, F_2, \dots, F_{12}$'. Owing to the consideration of big data, the objective remains in predicting the value 'Y_t' at time 't'.With this assumption, the sequence modeling network is then defined as a function '$Fun: F^{t+1} \rightarrow Y^{t+1}$', that generates the mapping function, '$Y_1, Y_2, \dots, Y_t = Fun\ (F_1, F_2, \dots, F_t)$' due to the big size of data. The main objective of marine weather forecast learning in the proposed work remains in identifying a function 'Fun' that minimizes the loss 'L' between the actual marine prediction outputs and the estimated marine predictions. Then, many-year long multivariate time series for big size of data employing E1 Nino big data dataset is mathematically formulated as given below.

$$Y(t) = F(t) + \Sigma(t) * W(t) \qquad (1)$$

$$\begin{bmatrix} Y_1(t) \\ Y_2(t) \\ \dots \\ Y_n(t) \end{bmatrix} = \begin{bmatrix} F_1(t) \\ F_2(t) \\ \dots \\ F_n(t) \end{bmatrix} \begin{bmatrix} \Sigma_{11}(t) & \Sigma_{12}(t) & \dots & \Sigma_{1n}(t) \\ \Sigma_{21}(t) & \Sigma_{22}(t) & \dots & \Sigma_{2n}(t) \\ \dots & \dots & \dots & \dots \\ \Sigma_{n1}(t) & \Sigma_{n2}(t) & \dots & \Sigma_{nn}(t) \end{bmatrix} \begin{bmatrix} W_1(t) \\ W_2(t) \\ \dots \\ W_n(t) \end{bmatrix} \qquad (2)$$

$$Y(t) = [Y_1(t) + Y_2(t) + \cdots + Y_n(t)] \qquad (3)$$

From the above equations (1), (2) and (3), 'n' represents the time series, with the vector '$F(t)$' and the matrix '$\Sigma(t)$' denotes the periodic functions with a period of one year, and finally, '$W(t)$' is supposed to be a vector zero-mean representing the disclosed seasonal patterns obtain from oceanographic and surface meteorological readings. Finally, from time series vector '$Y(t)$', '$Y_1(t)$' is the time series oceanographic and surface meteorological readings for the year '$Y1$', '$Y_2(t)$' is the time series oceanographic and surface meteorological readings for the year '$Y2$' and so on.

The binary classifier for marine weather forecast learning with time series data in perceptron involves a function that decides whether or not an input represented by vector of numbers belongs to particular class. This is mathematically expressed as given below.

$$Res = Y_i(t) = fun\ [W(t) * F_i] \qquad (4)$$

$$= fun\,[W_1(t)F_{i1} + W_2(t)F_{i2} + \cdots + W_n(t)F_{in}] \quad (5)$$

From the above equations (4) and (5), the actual output represented by vector to identify whether it belongs to a particular class or not is derived on the basis of function 'fun', input feature 'F_{in}' for a specific sample data 'i' initialized with weight 'W' respectively. The perceptron here makes feature selection on the basis of linear predictor function by integrating the set of weights with feature vector and updating the weight as given below.

$$W_i(t+1) = W_i(t) + LR * [D_i - Y_i(t)]F_{ij} \quad (6)$$

From the above equation (6), weight '$W_i(t+1)$' is updated according to the learning rate 'LR', desired output 'D_i' (i.e., feature selected), actual output '$Y_i(t)$' and prevailing weight '$W_i(t)+$' for all features 'F_{ij}'. The pseudo code representation of Linear Perceptron-based Feature Selection is given below.

Input: Dataset 'DS', Features '$F = F_1, F_2, \ldots, F_n$', Time '$t$'
Output: Robust and optimal feature selection 'FS'
Step 1: **Initialize** weight '$W = W_1, W_2, \ldots, W_n$', learning rate '$LR$'
Step 2: **Begin**
Step 3: **For** each Dataset 'DS' with Features 'F' and Time 't'
Step 4: **For** each sample 'i' in training Dataset 'DS'
Step 5: Obtain multivariate time series marine weather data as in equation (1), (2) and (3)
Step 6: Evaluate binary classifier for marine weather forecast as in equation (4) and (5)
Step 7: Obtain weight as in equation (6)
Step 8: **If** '$W * F + b > 0$'
Step 9: **Then** '$val(Res) = 1$'
Step 10: Feature 'FS' is selected
Step 11: Return feature selected 'FS'
Step 12: **End if**
Step 13: **If** '$W * F + b \leq 0$'
Step 14: **Then** '$val(Res) = 9$'
Step 15: Feature 'FS' is not selected
Step 16: **End if**
Step 17: **End for**
Step 18: **End for**
Step 19: **End**

Algorithm 1 Linear Perceptron-based Feature Selection

As given in the above Linear Perceptron-based Feature Selection for each marine weather dataset acquired as input consisting of twelve distinct features, the objective here remains in selecting the most robust and optimal features with maximum accuracy and minimum time. To achieve this objective, the samples present in the dataset is acquired as time series data and with this samples for twelve distinct features, binary classifier is deduced using Multivariate Perceptron function. Next, with the binary classifier output, the weights are updated according to the learning rate to obtain optimal features.

3.2 Kriging Ensembled eXtreme Gradient Boost Classification-based Marine Weather Forecasting

Upon successful completion of the robust and optimal feature selection, in this section, a classification-based marine weather forecasting model using Kriging Ensembled eXtreme Gradient Boost Classification is applied. The Kriging Ensembled eXtreme Gradient Boost Classification Process combines the weak learner (i.e., kriging regression) to form strong classifier. Figure 3 shows the block diagram of Kriging Ensembled eXtreme Gradient Boost Classification-based Marine Weather Forecasting.

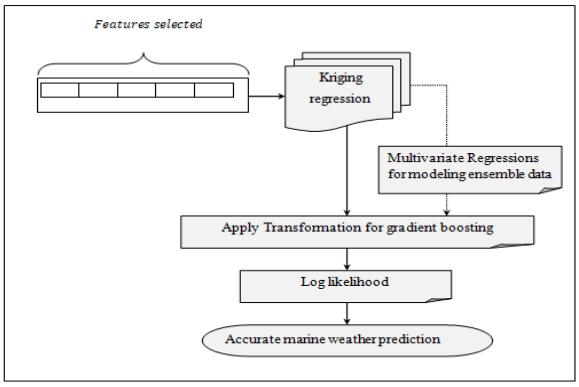

Figure 3 Block diagram of Kriging Ensemble eXtreme Gradient Boost Classification-based Marine Weather Forecasting

As shown in the above figure, Kriging regression is initially analyzed mathematically employing auxiliary predictors (i.e., feature selected) to estimate the kriging weights. Kriging regression analysis here involves a statistical process for estimating and analyzing several data (i.e., Big data) to identify relationship between weather dependent data (i.e., weather result) and one or more independent data of same data (e.g., relevant features).

The Kriging regression here integrates a regression model by simply kriging of regression residuals. Therefore, variogram residuals are first estimated and then, Simple Kriging (SK) is applied to the residual values to provide spatial prediction of the residuals. Kriging regression analysis

identifies variation of dependent data when any of factors in independent data gets changed. The Kriging regression analysis is then estimated as given below.

$$Z'_{KR}(FS) = Est'_{KR}(FS) + \sum_{\alpha=1}^{n(FS)} \lambda_{\alpha}^{KR}(FS) R(FS_{\alpha}) \quad (7)$$

From the above equation (7), ' $Est'_{KR}(FS)$ ' represents the regression estimate for the selected features 'FS' for kriging of regression 'KR' with residuals '$R(FS_{\alpha})$' (i.e., difference between the observed and predicted value) and 'α' value ranging between '1' and the size of the feature selected ' $n(FS)$ ' using an interpolation function ' λ ' respectively, therefore constituting weak learners. Gradient boosting in our work utilizes Kriging Regression. It provides a prediction model in the form of an ensemble of weak prediction models, and in our work it is Kriging Regression.

As Kriging Regression is the weak learner in our work, the resulting model is called as the Gradient Boosted Regression. Here, the residual of the current classifier forms as the input for the succeeding classifier on which the regressions are modeled and therefore it forms an additive model. The residuals are acquired gradually by the classifiers, with the purpose of encapsulating the maximum divergence within the data, by establishing learning rate to the classifiers for marine weather forecasting.

We employ a constraint of binary classifier regression, however in fact, Gradient Boosted Regression possess multivariate regressions. Hence, predictions are with reference to log however regression values range from integer to real values resulting in disparity. Therefore, a transformation is employed as given below.

$$T = \frac{\sum Res}{\sum [Prev\ Prob * (1 - Prev\ Prob)]} \quad (8)$$

From the above equation (8), the transformation for Gradient Boosted Regression is evaluated based on the residual 'Res' (i.e., representing the sum of residuals in that specific regression) and the sum of previous prediction probability ' $Prev\ Prob$ ' and same previous prediction probability '$1 - Prev\ Prob$' respectively. Finally, the log likelihood of predicted data is mathematically expressed as given below.

$$Log\ (likelihood\ [data]) = [Obs_i * \log(Pred) + (1 - Obs_i) * \log(1 - Pred)] \quad (9)$$

From the above equation (9), the log likelihood of data '$Log\ (likelihood\ [data])$', are arrived at on the basis of the observed data 'Obs_i' and the predicted data '$Pred$'. With this obtained results, the marine weather prediction is made in an accurate manner. The pseudo code representation of Kriging Ensembled eXtreme Gradient Boost Classification is given below.

Input: Dataset ' DS ', Features Selected ' $FS = FS_1, FS_2, \dots, FS_n$ '
Output: Accurate marine weather forecasting
Step 1: **Initialize** Step 2: **Begin** Step 3: **For** each Dataset 'DS' with Features Selected 'FS' Step 4: Formulate Kriging regression analysis as in equation (7) Step 5: Perform transformation as in equation (8) Step 6: Evaluate log likelihood of predicted data as in equation (9) Step 7: Return predicted results Step 8: **End for** Step 9: **End**

Algorithm 2 of Kriging Ensembled eXtreme Gradient Boost Classification

As given in the above algorithm, with the purpose of reducing the error rate involved in marine weather prediction, with the relevant features selected as input, first regression using Kriging is applied, therefore obtaining weak learners. Next, a transformation process is performed using Gradient Boosting due to various types of weak learners arrived at using the selected features. With this Gradient Boosting, several weak learners are combined or ensemble to evolve strong learner. Finally, accurate and precise classification of marine weather data is made by employing the log likelihood function. In this manner, accurate marine weather prediction is made with high accuracy and lesser time consumption, therefore reducing the error rate to a greater extent.

4. Experimental setup

The experimental settings of the proposed Perceptred-based Feature and Kriging Gradient Boost Classification (PF-KGBC) and two existing methods namely Spark-based fuzzy partitioning LSTM networks (SFPLN) [1] and Multilayer Convolutional Long and Short Term Memory (M-convLSTM) [2] are carried out using JAVA plugins and interfaces. In order to perform the experiment the

meteorological readings are taken from a series of buoys in the equatorial Pacific, i.e., E1 Nino Dataset (kaggle). To evaluate the performance of our method, we have used metrics such as prediction accuracy, prediction time and error rate with respect to number of weather data. To perform fair comparison, same amount of weather data possessing distinct sizes are obtained as input and comparisons are made with the state-of-the-art methods.

4.1 Dataset description

This E1 Nino dataset comprises of the readings obtained from oceanographic and surface meteorology via a series of buoys positioned throughout equatorial Pacific. Moreover, the data was obtained with Tropical Atmosphere Ocean (TAO) array that comprises of roughly 70 moored buoys spanning the equatorial Pacific that in turn measured oceanographic and surface meteorological variables with the purpose of detecting and predicting seasonal-to-inter annual climate changes occurring in the tropical regions. The variables or features present in this dataset are provided in table 1 with data obtained as 1980 for certain locations.

Some of the other pertinent data that was obtained in several locations are rainfall, solar radiation, current levels, and subsurface temperatures. In addition the data pertaining to latitude and longitude provided that the bouys were advanced throughput several locations. Also the values of the latitude were positioned within a degree from approximate location and the values of the longitude were positioned approximately five degrees off of the approximate location. Certain missing values are also present in the dataset.

Both the zonal and meridional winds oscillated between -10 m/s and 10 m/s and hence showed no linear relationship between the two wind positioning. The tropical Pacific values of relative humidity also ranged between 70% and 90%. Also the air temperature and sea surface temperature varied between 20 and 30 degrees Celcius and therefore showed positive linear relationship existing between the two types of temperature. Moreover, plots of other variables pertaining to meteorological data exhibited no linear relationship.

Not all buoys were potential to estimate values of currents, rainfall, and solar radiation. Hence, these values are said to be missing and according to the individual buoy are found to be dependent also. The quantity of data accessible is also dependent on the buoy, owing to the reason that specific buoys were employed preliminarily upon comparison with others.

4.2 Case 1: Impact of prediction accuracy

Prediction accuracy refers to the accurately predicted weather on specific data using the meteorological weather data. This is mathematically formulated as given below.

$$P_{acc} = \sum_{i=1}^{n} \frac{D_{pacc}}{D_i} * 100 \quad (10)$$

From the above equation (10), prediction accuracy 'P_{acc}' is measured based on the weather data involved in the simulation process for marine weather forecasting 'D_i' and the weather data accurately predicted 'D_{pacc}'. It is measured in terms of percentage (%). Table 1 shows the prediction accuracy comparison when using different predictors PF-KGBC, SFPLN [1] and M-convLSTM [2] on E1 Nino database.

Table 2 Tabulation of prediction accuracy using PF-KGBC, SFPLN [1] and M-convLSTM [2]

Weather data	Prediction accuracy (%)		
	PF-KGBC	SFPLN	M-convLSTM
5000	97.1	95.9	94.1
10000	94.35	90.45	87.35
15000	91.55	84.15	81.25
20000	86.35	81.35	77.45
25000	83.15	76.45	72.15
30000	80.25	73.35	68.55
35000	78.15	71.15	63.15
40000	75.35	67.55	60.25
45000	72	66	58.55
50000	70.15	63.35	54.15

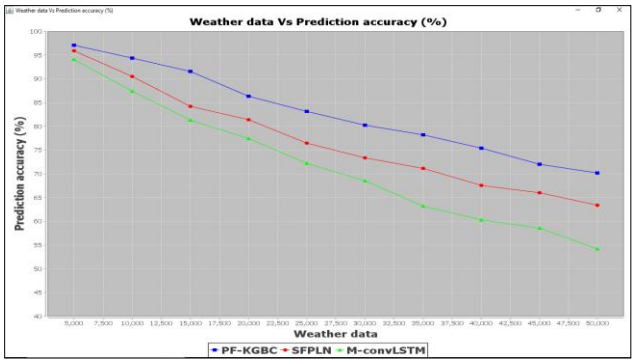

Figure 4 Graphical representation of prediction accuracy

Figure 4 graphically compare the results of prediction accuracy obtained from three different classification and prediction methods. X axis refers to the number of weather data ranging between 5000 and 50000 collected for simulations, consisting of both normal weather data and abnormality while measuring data. Y axis refers to the prediction accuracy obtained in terms of percentage (%). The prediction accuracy here is inversely proportional to the number of weather data provided as input. This is because the sample weather data provided as input includes both the normal and abnormal weather data and also due to the presence of irrelevant data, though feature selection being performed in all the three methods certain amount of irrelevancy are left to be unnoticed, therefore the results are said to be linear in results. However, simulations conducted with '5000' samples, shows that '4855' data were correctly measured using detected using PF-KGBC, '4795' samples using SFPLN [1] and '4705' images using M-convLSTM [2]. From this it is inferred that the prediction accuracy was found to be '97.1%' using PF-KGBC, '95.9%' and '94.1' using [1] and [2] respectively. This is because of the application of Linear Perceptron-based Feature Selection (LPFS) model. By applying this model, binary classifier for marine weather forecast learning using E1 Nino time series data with the aid of perceptron obtains a function. With this function, decision regarding represented by vector of numbers belongs to particular class or not is made. With this functional change, the set of weights with feature vector are integrated by means of weight updates according to the learning rate. In this way, the prediction accuracy using PF-KGBC is improved by 8% compared to [1] and 17% compared to [2] respectively.

4.3 Case 2: Impact of prediction time

The time consumed in predicting the marine weather using meteorological data is referred to as the prediction time. Weather prediction time is said to be one of the significant metrics to analyze the marine weather condition. This is due to the reason that minimum the weather prediction time faster the detection is said to be and more soon the positive measures can be taken. It is measured as given below.

$$P_{time} = \sum_{i=1}^{n} D_i * Time\ [Log\ (likelihood\ [data])] \quad (11)$$

From the above equation (11), prediction time 'P_{time}' is measured based on the weather data involved in the simulation process for marine weather forecasting 'D_i' and the time consumed in weather data prediction based on log likelihood function '$Time\ [Log\ (likelihood\ [data])]$'. It is measured in terms of milliseconds (ms). Table 3 shows the prediction time comparison using PF-KGBC, SFPLN [1] and M-convLSTM [2] with weather data acquired from E1 Nino dataset.

Table 3 Tabulation of prediction time using PF-KGBC, SFPLN [1] and M-convLSTM [2]

Weather data	Prediction time (ms)		
	PF-KGBC	SFPLN	M-convLSTM
5000	1750	2500	3250
10000	1835	2835	3550
15000	2015	3135	4035
20000	2235	3245	4135
25000	2415	3585	4285
30000	3125	4125	4735
35000	3255	4235	5015
40000	4135	4585	5185
45000	4185	5035	5855
50000	4235	5125	6015

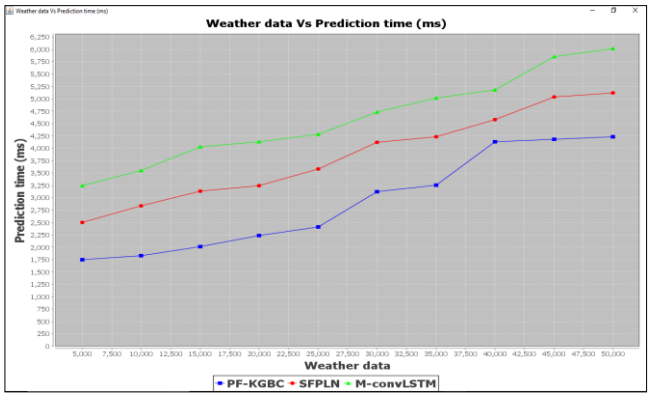

Figure 5 Graphical representation of prediction time

Figure 5 given above shows the graphical image of the weather detection time with number of weather data ranging from 5000 to 50000 as input collected from E1 Nino dataset. From the figure it is inferred that the weather detection time is directly proportional to the number of weather data. This is because increasing the number of weather data, the time consumed in obtaining optimal feature selection increases and therefore the weather detection time also increases. However, from the simulation of '5000' number of weather data the time consumed in obtaining optimal features using PF-KGBC is '$0.35ms$', '$0.50ms$' using SFPLN [1] and '$0.65ms$' using M-convLSTM [2]. With this the weather prediction time using PF-KGBC was observed to be '$1750ms$', '$2500ms$' and '$3250ms$' using [1] and [2] respectively. From the simulations it is inferred that the weather prediction time using PF-KGBC is comparatively lesser than [1] and [2]. This is because of the application of Linear Perceptron-based Feature Selection algorithm. By applying this algorithm, the weather data samples present in the dataset is obtained in the form of time series data and with weather data samples for twelve distinct features, binary classifier was applied using Multivariate Perceptron function. With the obtained results, the weights were updated based on the learning rate to acquire optimal features. This in turn reduced the weather prediction time using PF-KGBC method by 26% compared to [1] and 38% compared to [2] respectively.

4.4 Case 3: Impact of Error rate

Finally error rate is measured to analyze the deviations observed during the marine weather forecasting. This is mathematically formulated as given below.

$$ER = MAE = \frac{1}{n}\sum_{i=1}^{n}(AV_i - PV_i) * 100 \quad (12)$$

From the above equation (12), error rate 'ER' is measured on the basis of mean absolute error 'MAE' arrived at on the basis of the actual value 'AV_i' and the predicted value 'PV_i' for 'n' samples respectively. It is measured in terms of percentage (%). Table 4 shows the error rate comparative analysis using PF-KGBC, SFPLN [1] and M-convLSTM [2] with weather data acquired from E1 Nino dataset.

Table 4 Tabulation of error rate using PF-KGBC, SFPLN [1] and M-convLSTM [2]

Weather data	Error rate (%)		
	PF-KGBC	SFPLN	M-convLSTM
5000	1.6	2.8	4.6
10000	1.8	3.2	5
15000	2.3	3.5	5.5
20000	2.5	3.9	5.8
25000	2.8	4	6
30000	3	4.3	6.5
35000	3.2	4.5	6.9
40000	3.5	5	7.5
45000	4	5.5	8
50000	4.5	7	9

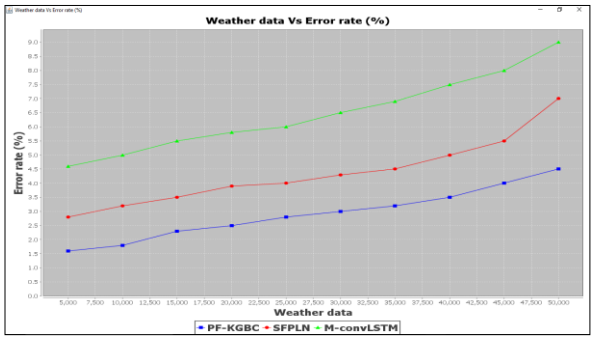

Figure 6 Graphical representation of error rate

Finally, the figurative representation of error rate with respect to 50000 distinct weather data is shown in the above figure. From the figure it is inferred that with small amount of weather data, the error rate was less and increasing the size of weather data resulted in the increase in the error

rate, therefore the error rate is said to be directly proportional to the number of weather data provided as input for simulation. However, the error rate is significantly reduced by applying PF-KGBC method upon comparison with SFPLN [1] and M-convLSTM [2]. With simulations performed for sample of 5000 weather data, with the actual value being '4395', the predicted value using the three methods were observed to be '4855', '4795' and '4705', therefore the error rate being, '1.6%', '2.8%' and '4.6%' respectively. The reason behind the minimization of error rate using PF-KGBC method was owing to the Kriging Ensembled eXtreme Gradient Boost Classification-based Marine Weather Forecasting. Here, first Kriging regression analysis was made with the features selected, followed by which gradient boosting was applied for ensemble results. As a result, several weak learners are combined or ensemble to evolve strong learner, therefore reducing the error rate using PF-KGBC method by 34% compared to [1] and 56% compared to [2].

5. Conclusion

This paper proposed a novel method called Perceptred-based Feature and Kriging Gradient Boost Classification (PF-KGBC) with Big Data for marine weather forecasting, which contains two stages, i.e., feature selection and classification. In the first stage, robust and optimal features are selected by not only eliminating the irrelevant and redundant features using binary classifier function, optimal features are selected using Perceptron function. Finally, to predict the marine weather data, Kriging Ensembled eXtreme Gradient Boost Classification is applied that possess the advantage of ensembling the weak classified results to strong classifier log likelihood of data, therefore reducing the error rate. The performance of proposed PF-KGBC method is determined in terms of weather prediction accuracy, prediction time and error rate with respect to different huge numbers of weather data compared with two traditional methods. The simulation result demonstrates that proposed PF-KGBC method gives better performance when compared to state-of-the-art works.

References

[1] Jiabao Wen, Jiachen Yang, Bin Jiang, Houbing Song, and Huihui Wang, "Big Data Driven Marine Environment Information Forecasting: A Time Series Prediction Network", IEEE Transactions on Fuzzy Systems, Volume 29, Issue 1, January 2021, Pages 4-18 [Spark-based fuzzy partitioning LSTM networks (SFPLN)]

[2] Kun Zhang, Xupu Geng, and Xiao-Hai Yan, "Prediction of 3-D Ocean Temperature by Multilayer Convolutional LSTM", IEEE Geoscience and Remote Sensing Letters, Volume 17, Issue 8, August 2020, Pages 1303-1307

[3] Partha Pratim Sarkar, Prashanth Janardhan and Parthajit Roy, "Prediction of sea surface temperatures using deep learning neural networks", SN Applied Sciences, Springer, Volume 2, Issue 1458, 2020, Pages 1-15

[4] Hedi Kanarik, Laura Tuomi, Jan-Victor Bj, orkqvist, Tuomas Karna1, "Improving Baltic Sea wave forecasts using modelled surface currents", Ocean Dynamics, Springer, Apr 2021

[5] Ari Yair Barrera-Animas, Lukumon O. Oyedele, Muhammad Bilal, Taofeek Dolapo Akinosho, Juan Manuel Davila Delgado, Lukman Adewale Akanbi, "Rainfall prediction: A comparative analysis of modern machine learning algorithms for time-series forecasting", Machine Learning with Applications, Elsevier, Oct 2021

[6] Krishna Mishra, M Sharma, M Mohapatra, "Performance of numerical weather prediction models in predicting track of recurving cyclone Vayu over Arabian Sea during June 2019", Journal of Earth System Science, Springer, Nov 2020

[7] Chuang Zhang, Ming Wu, Jinyu Chen, Kaiyan Chen, Chi Zhang, Chao Xie, Bin Huang, Zichen He, "Weather Visibility Prediction Based on Multimodal Fusion", IEEE Access, May 2019

[8] D. Menaka and Sabitha Gauni, "Prediction of Dominant Ocean Parameters for Sustainable Marine Environment", IEEE Access, Oct 2021

[9] Sue Ellen Haupt and Branko Kosovi´c, "Variable Generation Power Forecasting as a Big Data Problem", IEEE Transactions on Sustainable Energy, Vol. 8, No. 2, Apr 2017

[10] L. Apicella, S. Puca, M. Lagasio, A. N. Meroni, M. Milelli, N. Vela, V. Garbero, L. Ferraris, A. Parodi, "The predictive capacity of the high resolution weather research and forecasting model: a year-long verification over Italy", Bulletin of Atmospheric Science and Technology, Springer, Apr 2021

[11] Clive E. Dorman, Andrey A. Grachev, Ismail Gultepe, Harindra J. S. Fernando, "Toward Improving Coastal-Fog Prediction (C-FOG)", Boundary-Layer Meteorology, Springer, Nov 2021

[12] Pradeep Hewage, Ardhendu Behera, Marcello Trovati, Ella Pereira, Morteza Ghahremani, Francesco Palmieri, Yonghuai Liu, "Temporal convolutional neural (TCN) network for an effective weather forecasting using time-series data from the local weather station", Soft Computing, Springer, Apr 2020

[13] Vinay Kellengere Shankarnarayan, Hombaliah Ramakrishna, "Comparative study of three stochastic future weather forecast approaches: a case study", Data Science and Management, KeAi, Chinese Roots Global Impact, Jul 2021

[14] K.U. Jaseena, Binsu C. Kovoor, "Deterministic weather forecasting models based on intelligent predictors: A survey", Journal of King Saud University – Computer and Information Sciences, Elsevier, Sep 2020

[15] Ashutosh Kumar Tanvir Islam, Yoshihide Sekimoto, Chris Mattmann, Brian Wilson, "Convcast: An embedded convolutional LSTM based architecture for precipitation nowcasting using satellite data", PLOS ONE | https://doi.org/10.1371/journal.pone.0230114 March 11, 2020

[16] Pradeep Hewage, Marcello Trovati, Ella Pereira, Ardhendu Behera, "Deep learning-based effective fine-grained weather forecasting model", Pattern Analysis and Applications, Springer, Jun 2020

[17] Bartosz Skobiej, Arto Niemi, "Validation of copula-based weather generator for maintenance model of offshore wind farm", WMU Journal of Maritime Affairs, Springer, Oct 2021

[18] Jan Kleine Deters, Rasa Zalakeviciute, Mario Gonzalez, and Yves Rybarczyk, "Modeling PM2.5 Urban Pollution Using Machine Learning and Selected Meteorological Parameters", Journal of Electrical and Computer Engineering, Hindawi, Jun 2017

[19] Shaoqing Zhang, Guijun Han, Yuanfu Xie, Juan Jose Ruiz, "Data Assimilation in Numerical Weather and Climate Models", Advances in Meteorology, Hindawi, Jun 2015

[20] Kristoko Dwi Hartomo and Yessica Nataliani, "A new model for learning-based forecasting procedure by combining k-means clustering and time series forecasting algorithms", Peer J Computer Science, Jun 2021