2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)
Banff Center, Banff, Canada, October 5-8, 2017

Ensemble Learning for Forecasting Main Meteorological Parameters

Petros Karvelis
Laboratory of Knowledge and Intelligent Computing
Department of Computer Engineering Technological
Educational Institute of Epirus, Arta, Greece

Stavros Kolios
Computer Technology Institute & Press "Diophantus"
Patras, Greece

George Georgoulas
Robotics Group, Control Engineering Division of the
Department of Computer, Electrical and Space
Engineering, Lulea University of Technology, Lulea
97187, Sweden

Chrysostomos Stylios
Laboratory of Knowledge and Intelligent Computing
Department of Computer Engineering Technological
Educational Institute of Epirus, Arta, Greece

***Abstract*—The significant role of predicting weather conditions in daily life, the new era of innovative machine learning approaches along with the availability of high volumes of data and high computer performance capabilities, creates increasing perspectives for novel improved short-range forecasting of main meteorological parameters. Among the various algorithms for forecasting parameters, ensemble learning approaches are able to generate simple models which provide accurate predictions for regression problems. The advantage of ensembles with respect to single models is that they perform remarkably well for a variety of problems. The main aim of this ongoing research is to provide some preliminary assessment of the applicability of ensemble learning for wind speed forecasting. In this work, forecasting results of a single and two ensemble models are presented and compared.**

Keywords— forecasting weather conditions; ensemble classifier; random forests, gradient boosted trees;

I. INTRODUCTION

The meteorological parameters are considered critical factors affecting many phenomena in the atmosphere, the climate, the weather and the physical environment on the Earth surface. Their variations at different spatiotemporal scales can significantly affect daily activities. Their extreme high/low values (extreme weather conditions) highly influence everyday life and especially the commercial and the transportation sector. Moreover, extreme weather can affect the safety of both inland and sea transportation, often causing many serious accidents with human losses and huge impact in the environment.

The frequency of occurrence of weather extremes are not the same worldwide thus robust and modern methodologies are needed to study and warn about such conditions. Methodologies for the estimation of the meteorological parameters, can offer useful informational background for warning about extreme weather phenomena, sustainable environmental management, safety in transportation and energy production [1], [2]. As a consequence, the continuous recording of the meteorological parameters is important. However, more important could be considered the accurate forecasting of these parameters.

In the literature, there is a number of studies that try to analyze time-series of different meteorological parameters. Some studies classify the weather types and try to estimate a potential climate change [3]-[5] while other studies try to analyze extreme weather events [6]-[9]. A wide range of methods have been used for long-term/short-term forecasting and the evolution of these parameters [10]-[15]. Nevertheless, the complex nature of these meteorological parameters can seriously influence prediction accuracy of any forecasting algorithm especially when the outcomes of a study are used in a different geographical area [16].

Ensemble learning offers significant advantages to single models in a way that makes them perform remarkably well for a variety of problems. Here, it is presented a preliminary assessment of the applicability of ensemble learning methods for the prediction of the wind speed (WS). The Multiple Linear Regression (MLR) [17] model is compared against two ensemble learning methods: the Gradient Boosted Trees (GBT) [17] and the Random Forests (RFs) [19]. For this task, three-months of data sets with temporal resolution of 15 minutes are used. The time series come from a coastal location of Greece (Corfu). The aim of this study is to develop an integrated application for short-range forecasting of basic meteorological parameters near coastlines in order to improve the forecasting of weather extremes and enhance the sea safety at coastal regions.

Section II briefly presents the applied methods and section III describes the available data that are used. Section IV presents the achieved results and the accuracy evaluation of the proposed methodological scheme and section V concludes the paper.

II. METHODS

Regression is a machine learning problem considering a set of training examples e.g. n pairs of data $\{x_i, y_i \mid 1 \le i \le n\}$ where $x_i \in \Re^m$ are the m regressors and $y_i \in \Re$ are the

corresponding prediction values. Then one can define regression, as the problem of finding a transformation $X_{n\times m} \xrightarrow{f} Y_n$, which minimizes the error between the actual value and the value predicted from the transformation. In this study, the available regressors are: Temperature, Dew point temperature, Humidity, Wind direction, Pressure, Precipitation, which are used to predict the WS parameter. Three models namely the MLR, the RFs and the GBTs are used.

A. Multiple Linear Regression

MLR is the simplest form of linear regression analysis. MLR, is commonly used to express the relationship between two or more independent parameters.

The training input data can be described by an augmented matrix X of size $n\times(m+1)$ where x_{ij} ($2 \le j \le m$) denotes the values of j -th parameter for the i -th observation:

$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1m} \\ 1 & x_{21} & \cdots & x_{2m} \\ 1 & \vdots & \cdots & \vdots \\ 1 & x_{n1} & \cdots & x_{nm} \end{pmatrix}. \quad (1)$$

The column of ones allows models to have an offset (not pass through the origin). A response vector y of size n where y_i denotes the value for the i -th observation of the parameter that needs to be predicted:

$$y = \begin{pmatrix} y_1 & y_2 & \cdots & y_n \end{pmatrix}^T. \quad (2)$$

Then a linear model can be defined as:

$$y = Xb + \varepsilon, \quad (3)$$

where $b \in \Re^m$ are the regression coefficients and $\varepsilon \in \Re^n$ are the residual errors. The regression coefficients can be easily found by a least square method for which the total residual error is minimized:

$$\beta = (\mathrm{X}^T\mathrm{X})^{-1} X^T y. \quad (4)$$

B. Gradient Boosted Trees

Decision Trees (DTs) is one of the most widely applied data mining method for classification and/or regression [20]. Their popularity stems from a number of advantages e.g. they can handle irrelevant and redundant parameters, they can handle continuous, discrete and categorical variables together, scaling of the variables does not matter and finally the decision process can be traced as a sequence of simple choices and training is reasonably fast.

DTs are able to divide the feature space into a number (J) of regions $R_j, 1 \le j \le J$ and then fit a specific model in each one of them [21]. In the case of regression, the DT model assigns a constant value $v_j, 1 \le j \le J$ to each region as:

$$if\,(x \in R_j)\ \text{then}\ y = v_j. \quad (5)$$

Thus, a Regression Tree (RT) can be defined as:

$$RT(x) = \sum_{i=1}^{J} I(x \in R_j) \cdot v_j, \quad (6)$$

where $I(x \in R_j) = \begin{cases} 1, & if\ x \in R_j \\ 0, & otherwise \end{cases}$.

Furthermore, tree based methods can produce high accurate predictions if grouped together in the form of an ensemble. One such approach uses the boosting principle, in order to create gradient boosted machines [22], [23]. The approach starts by building a simple model and then stage wise adds models that aim to explain observations that are modeled poorly by the existing trees of the ensemble ending up with a model of the form :

$$f(x) = \sum_{i=1}^{M} RT_i(x), \quad (7)$$

where M is the number of trees of the ensemble and RT_i is the i -th member of the ensemble [24].

C. Random Forests

RFs are another example of ensemble learning paradigm. These Forests comprise by a large set of DTs that function together in order to predict the value of a variable [19]. Each DT of the forest is created using a different bootstrap sample from the training set and each node of the DT is split using a random feature.

More specifically, for each node of the DT a subset S with a number of features F_S is selected ($F_S < m$). The best feature among the possible m features is selected for the node to be split.

III. DATA

The experimental data set consists of 15-minute records of seven parameters that are presented at Table I. They have been acquired by the global weather service Weather Underground[1], using an automated procedure, which has been developed to communicate with the relative service and to collect automatically all the available measurements in JavaScript Object Notation (json) format.

All the obtained datasets are thoroughly checked for possible errors and missing data before further analyzed. The collection of the data spans from January 1, 2017 till March 31, 2017. The dataset comes from a coastal meteorological station because our long-term scope of the study is to develop a stand-alone

[1] http://www.wunderground.com

application to enhance the sea safety during extreme weather conditions along the costal line.

TABLE I. BASIC METEOROLOGICAL PARAMETERS WHICH ARE USED IN THIS STUDY

Parameter	Units	Temporal Resolution
Temperature (T)	°C	15-min
Dew Point Temperature (DPT)	°C	15-min
Humidity (H)	%	15-min
Wind Direction (WD)	0°—360°	15-min
Pressure (PR)	hPa	15-min
Precipitation (PC)	mm	15-min
Wind Speed (WS)	Km/h	15-min

Each forecasting model uses the current and the previous two values of wind Temperature (T), Dew Point Temperature (DPT), Humidity (H), Wind Direction (WD), Pressure (PR), Precipitation (PC) and Wind Speed (WS) resulting in a vector of 21 values. Finally, each model predicts the value of the WS for the next 15 mins. In order to determine the value of the time lag we did a correlation analysis for the output parameter (WS) using the training data.

IV. RESULTS

The original data set is divided into a training set (80%) and a test set (20%) in order to measure the accuracy of the models. For the training set data from 72 consecutive days are used and the rest 9 consecutive days are used as a test set. The performance is quantified using the Mean Absolute Error (MAE), and the R^2 statistic, which are two popular yet not the only available measures [25].

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{8}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{9}$$

where y_i is the value of the i-th observation, $\hat{y}_i$ is the predicted value for the i-th observation and $\bar{y}$ is the mean value of the observations.

In order to validate our results, the Naive Forecasting 1 (NF1) [26] is also used for the prediction of WS. The NF1 model uses the most recent observation as the prediction for the next one.

The performance measures of the models used in this study are summarized in Table II. The best achieved values are depicted in bold.

TABLE II. THE PERFORMANCE OF THE MODELS USED FOR FORECASTING THE WIND PARAMETER

	NF1	MLR	RF	GBT
MAE	1.231	1.177	1.224	**1.167**
R^2	0.611	0.717	0.692	**0.718**

Fig. 1, presents a portion of the predicted WS for the Corfu station using GBTs.

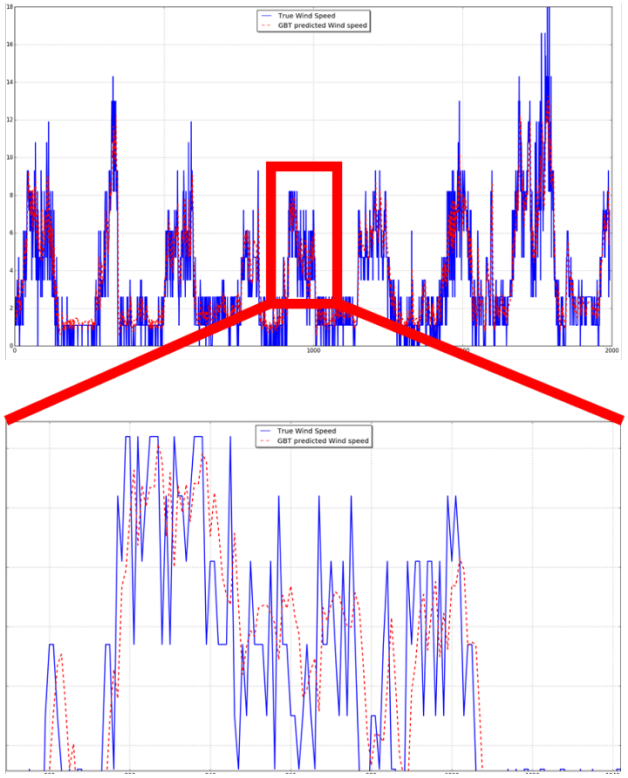

Fig. 1. The predicted WS with GBTs (the observed WS is displayed with blue line and with red line the predicted value).

Here it should be noted that sometimes a single prediction value may not suffice for some application fields. To this end, quantile regression constructs prediction intervals for new observations [27]. This type of regression is especially useful in applications where the extreme values are studied, such as environmental studies where upper quantiles are critical from a public health and safety perspective. Indeed, such approach (levels and interval of significance) can allow to provide reliable information for prediction considering the high variability and the randomness of the extreme values in WS. Fig. 2, displays the predicted and the real value of the wind for 90% of prediction intervals.

V. CONCLUSIONS

This work presented our preliminary results for the prediction of WS using previous values of the speed as well as other meteorological parameters. The results suggest that the use of GBTs can be beneficial for the prediction of meteorological

parameters. However more extensive experimentation is needed before deploying the method for industrial use. A successful implementation of a wind forecasting method can be beneficial in light of the general trend of building wind turbine farms for increasing the quota of green energy productions as well as for increasing safety in marine transportation.

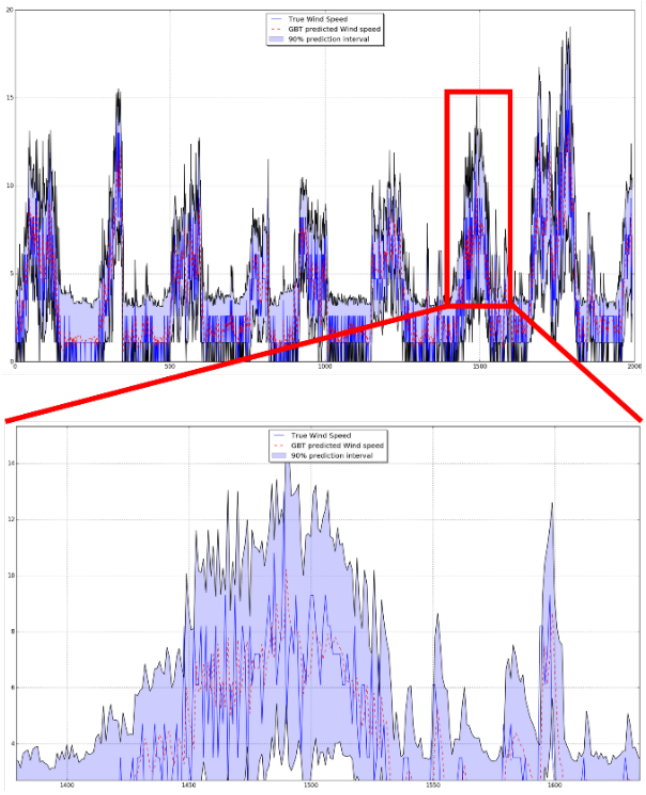

Fig. 2. Wind 90% prediction interval for the Corfu Station,

ACKNOWLEDGMENT

This study is supported by the National funds allocated by the Greek General Secretarial of Research and Development project 2006SE01330025 as continuation of FP7–PEOPLE–IAPP–2009, Grant Agreement No. 251589, Acronym: SAIL and by "LINCOLN" (Lean Innovative Connected Vessels) Project (www.lincolnproject.eu) Horizon 2020 research and innovation program (Grant Agreement: 727982).

REFERENCES

[1] IPCC, 2012. Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation. A Special Report of Working Groups I and II of the Intergovernmental Panel on Climate Change (IPCC), Cambridge University Press.

[2] S. Pfahl, "Characterising the relationship between weather extremes in Europe and synoptic circulation features," Nat. Hazards Earth Syst. Sci., vol. 14, pp. 1461–1475, 2014.

[3] A. R. Naik and S. K. Pathan, "Weather classification and forecasting using back propagation feed-forward neural network," International journal of scientific and research publications, vol. 2, no. 12, pp. 2250-3153, 2012.

[4] F. Olaiya and A. B. Adeyemo, "Application of data mining techniques in weather prediction and climate change studies," International Journal of Information Engineering and Electronic Business, vol. 1, pp. 51-59, 2012.

[5] M. Saha, P. Mitra and A. Chakraborty Fuzzy Clustering Based ensemble approach to predicting Indian Monsoon, Advances in Meteorology, 2015.

[6] K. Goubanova and L. Li, "Extremes in temperature and precipitation around the Mediterranean basin in an ensemble of future climate scenario simulations," Global and Planetary Changes, vol. 57, pp. 27-42 2007.

[7] A. Kalimeris, D. Founda, C. Giannakopoulos and F. Pierros, "Long term precipitation variability in the Ionian islands (Central Mediterranean): Climatic signal analysis and future projections," Theoretical and Applied Climatology, vol. 109, pp. 51–72, 2011.

[8] A. F. Karagiannidis, T. Karacostas, P. Maheras and T. Makrogiannis "Climatological aspects of extreme precipitation in Europe, related to mid-latitude cyclonic systems," Theoretical and Applied Climatology, vol. 107, pp. 165-174, 2012.

[9] M. S. Varfi, T. S. Karacostas, T. J. Makrogiannis and A. A. Flocas "Characteristics of the extreme warm and cold days over Greece," Advances in Geosciences, vol. 20, pp. 45-50, 2009.

[10] S. A. P. Kani and M. M. Ardehali "Very short-term wind speed prediction: A new artificial neural network-Markov chain model," Energy Conservation and Management, vol. 52, no. 1, pp. 738-745, 2011.

[11] Z. Guo, D. Chi, J. Wu and W. Zhang "A new wind speed forecasting strategy based on the chaotic time series modelling technique and the Apriori algorithm," Energy Conservation and Management, vol. 84, pp. 140-151, 2014.

[12] A. Pierre and V. Monbet "Markov-switching autoregressive models for wind time series," Environmental Modelling and Software, vol. 30, pp. 92-101, 2012.

[13] S. Chattopadhyay, D. Jhajharia G. Chattopadhyay, "Univariate modelling of monthly maximum temperature time series over northeast India: neural network versus Yule-Walker equation based approach," Meteorological Applications, vol. 18, pp. 70-82, 2011.

[14] F. Almonacid, P. Perez-Higueras, P. Rodrigo and L. Hontoria, "Generation of ambient temperature hourly time series for some Spanish locations by artificial neural networks," Renewable Energy, vol. 51, pp. 285-291, 2013.

[15] K. Abhishek, M. P. Sing, S. Ghosh and A. Abhishek, "Weather forecasting model using Artificial Neural Network," Procedia Technology, vol. 2, pp. 311-318, 2012.

[16] G. Georgoulas, P. Karvelis, S. Kolios and C. Stylios, "Examining nominal and ordinal classifiers for forecasting wind speed," in Proceedinds of 8th IEEE International Conference on Intelligent Systems IS'16. 3-5 Septeber 2016, Sofia, Bulgaria. pp. 504-509.

[17] G. Seber and A. Lee, Linear Regression Analysis, 2nd Edition, Wiley & Sons, New Jersey, 2003.

[18] G. Ridgeway, The state of boosting. Computing Science and Statistics, 172-181, 1999.

[19] L. Breiman, "Random forests," Machine Learning, vol. 45, pp. 5–32, 2001.

[20] G. Seni and J. F. Elder. Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions. Synthesis Lectures on Data Mining and Knowledge Discovery. Morgan & Claypool Publishers, 2010.

[21] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, S. Y. Philip et al., "Top 10 algorithms in data mining," Knowledge and information systems, vol. 14, no. 1, pp. 1-37, 2008.

[22] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," Annals of statistics, pp. 1189-1232, 2001.

[23] K. Ludmila, Combining Pattern Classifiers, 2nd Edition, Wiley, 2014.

[24] J. Friedman, T. Hastie and R. Tibshirani, The elements of statistical learning, volume 1. 2nd Edition, 2009.

[25] A. Saxena, J. Celaya, B. Saha, S. Saha and K. Goebel, "Metrics for offline evaluation of prognostic performance," International Journal of Prognostics and Health Management, vol. 1, no. 1, pp. 4-23, 2010.

[26] S. Makridakis, S. Wheelwright and R. Hyndman, Forecasting Methods and Applications, 3rd Edition, Wiley, 2012.

[27] N. Meinshausen, "Quantile regression forests," Journal Machine Learning Research, vol. 7, pp. 983–999, 2006.