An Intelligent System for Filling of Missing Values in Weather Data

Maqsood Ali Solangi[1], Dr. Ghulam Ali Mallah[2], Shagufta Naz[3], Jamil Ahmed Chandio[2], Muhammad Bux Soomro[1]

Department of Computer Science, Shaheed Zulfikar Ali Bhutto Institute of Science and $Technology^1$
Larkana, Pakistan
Department of Computer Science, Shah Abdul Latif University Khairpur $Mir's^2$
Department of Basic Sciences & Related Studies, Benazir Bhutto Shaheed University of Technology and Skill Development (BBSUTSD) Khairpur $Mir's^3$

Abstract
Recently Machine Learning has been considered as one of the active research areas of Computer Science. The various Artificial Intelligence techniques are used to solve the classification problems of environmental sciences, biological sciences, and medical sciences etc. Due to the heterogynous and malfunctioning weather sensors a considerable amount of noisy data with missing is generated, which is alarming situation for weather prediction stockholders. Filling of these missing values with proper method is really one of the significant problems. The data must be cleaned before applying prediction model to collect more precise & accurate results. In order to solve all above stated problems, this research proposes a novel weather forecasting system which consists upon two steps. The first step will prepare data by reducing the noise; whereas a decision model is constructed at second step using regression algorithm. The Confusion Matrix will be used to evaluation the proposed classifier.

Keywords
Machine Learning, Missing Values, data cleaning, Weather Prediction

I. INTRODUCTION

The prediction of weather data is one active research area of computer science and data of epidemiology of diseases, earthquakes, heatwaves, and others are providing assistance to formulate the proper policies. Since noisy data with missing values need to be cleaned with proper methods would boost the prediction model. For instance, heterogynous and malfunctioning of weather senor data may become cause of confusion when it sends improper data. Particularly missing values would be considered a major problem in weather data because noisy data would produce noisy results. The careful investigation is really important for Noisy data generated by the weather senor may consist unpredictable and produce inaccurate results. The efficiency of Weather stations has been improved however, the failures still occur and techniques are required to fill gaps in data sequences in order to use them as inputs in weather or energy models. Many approaches have been followed by researchers. In weather data Co-relation Coefficient, Root Mean Squared Error, aggregation, and other statistical measures may not be filled the missing values because time series data must be filled with proper binning techniques to fill the missing values to generate the results accurately. The methodology of this paper comprises over three stages, in first step, data preparation technique is proposed where rather than elimination of irrelevant and inconsistent data, we replace the quantiles with proposed binning techniques as described in methodology section. In second step we use time series regression to construct the decision model to find out the hidden patterns of data and the evolution of classifier is done by using confusion matrix. We have used open source datasets, downloaded from UCI Machine learning repository for weather data.

This article is divided into five sections where introduction is described at section one, whereas section two presents the related works, meanwhile methodology is defined in section three. The results are elaborated at section four whereas conclusion and discussion is discussed at section five.

II. LITERATURE BASED BACK GROUND OF THE PROBLEM

An approach [1] was proposed by using machine learning techniques on a GIS platform to assess the air pollution. NOAA, linear regression machine learning techniques were used to estimate an association among the variables and an open source GIS system was used to visualized the air pollutant hazards were show that high density population greater than 1200 number of sectors were reported as 1797 for plant-mercury emissions under the age of 5 years and below, 800-1200 for medium density population was recorded as 15761 and 400-800 number of tracts on a GIS dependent shows 11665 Barry Plant-mercury emissions were included in the low-density population.

A visualization of weather forecasting system [2] was proposed by combining Artificial Neural networks and GIS based techniques. Artificial technique was used to measure the quantification of weather variables to predict the minimum and maximum temperature levels in Jordan. The

Manuscript received September 5, 2023
Manuscript revised September 20, 2023
https://doi.org/10.22937/IJCSNS.2023.23.9.12

results show that the minimum (High, Low) average temperature levels on a zone from 1979 to 1988 and 1999 to 2008-2018. Rainfall variable may vary the temperature in all zones else the temperature may be recorded increased.

A hybrid intelligent approach [3] was proposed to visualize the forest fire mode on GIS map by using machine learning techniques. PSO-NF, Random forest and support vector machine AI based techniques were used to resolve the complex problem of forest fire system the best accuracy was measured as 91%

different standardized mean error and the results show that the performance of Kriging techniques was better than non-geostatistical interpolation techniques.

An intelligent System [8] was proposed to predict the surrounding temperature or humidity in an environment. Linear regression machine learning techniques was used to measure the temperature sensing data based the input variables and the result shows the accuracy and lowest possible time of the system and the total score was tested as 100%.

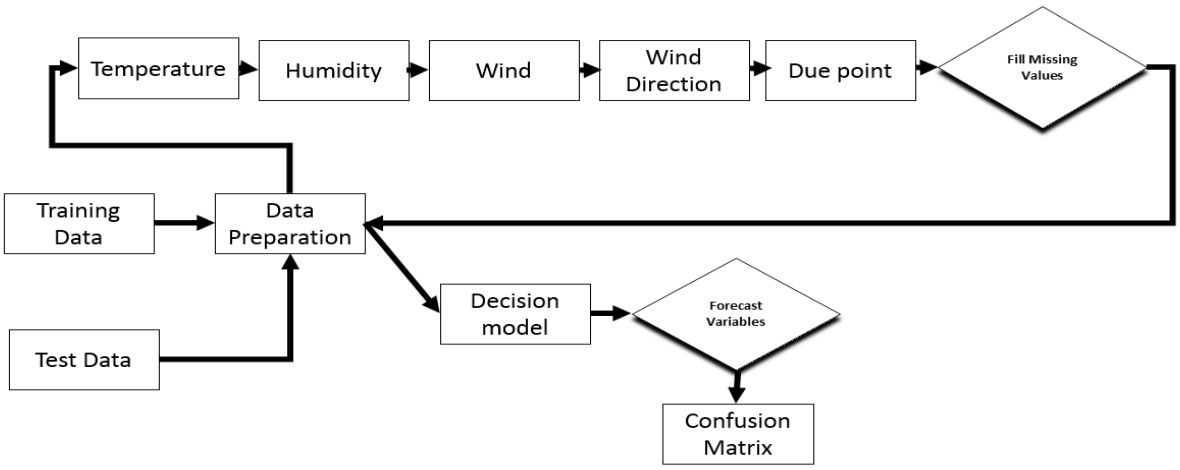

Figure 1: An Intelligent System for Filling of Missing Values in Weather Data Workflow

Prediction of humidity sensor was analyzed [4] to enhance the health life testing using artificial intelligent techniques among the variables on the basis of minimum error and maximum accuracy. DHT11, ANNs, FIS and ANFIS was used for analyze to choose the best techniques for results. The results were shown in three different techniques that show the accuracies of the comparison, the best minimum average error 2.48% and efficient accuracy 97.57% was calculated by using ANFIS technique.

A Hybrid system was proposed [5] to get the real-time soil in agriculture land based on GSM to predict the soil moisture content (MC) hourly. Fuzzy logic was used to construct the decision model for irrigation variables and the results show that the RSE was recorded as 0.985 and applied on 11 different soil types

A system was proposed [6] to visualize the moisture data on GIS platform. GIS classification techniques were used to classify four distinct categories to find out the average moisture index of the Rock and soil. The results show that all areas need different level of water to irrigate properly.

A comparative study [7] of nine GIS spatial interpolation techniques was conducted for measure the temperature of heat- related health risk in subtropical city. Interpolation techniques were used to validate the data by comparing

An approach [9] was proposed to visualize the water temperature by merging GIS technology and Artificial Neural Network (ANN). Multiple linear regression machine learning techniques was used to estimate the temperature by using ANN models. The results show that the R2 was recorded as 0.79 with MSE as 0.29 based on the different input parameters

III. Proposed Material And Methods

The proposed methodology offers predictive mining to fill the missing values of forecast variables. There are two steps in our proposed methodology. Step one describes data preparation method which consists over sever sub-steps such as recording of temperature, humidity, wind, wind direction and due point. Whereas a careful observation is made to fill the missing values by binning technique. The second step is constructing a decision model to explore the insight knowledge of associated variables and validation was done by confusion matrix.

A) PRE-PROCESSING:

Dataset definition:

Data preparation is one of the challenging task because noisy data would produce inaccurate results and in our use case missing values are one of difficult task since the variables of weather data could not be aggregated. The temperature is divided into two classes. Class one shows the high temperature and class two is with low temperature. Humidity is also divided into two distinct categories such as upper level and lower level. Wind speed is recorded as low and high whereas wind direction is divided into four classes along with angle from with it is blooming. Due point is recorded as present and absent. All the above stated variables are associated which are to be mapped consisting upon the administrative areas such as districts.

FILLING OF MISSING VALUES:

Consider that there exist anomalies in data, we deal dual approaches to fill the missing values, in first dataset we record variables quantities by using binning technique where three neighborhood values are replaced with mean values and secondly, we use mean as well as standard deviation to fill the missing values.

B) DECISION MODEL:

By considering the nature of time series data, time series regression machine learning technique is applied to construct the decision model and to co-relate the variables for forecasting. The standard statistical method ARIMA was selected because of having auto regressive moving average build effective models to cope the situation such as weather data is constantly observed and it needs online forecasting system to measure the activities.

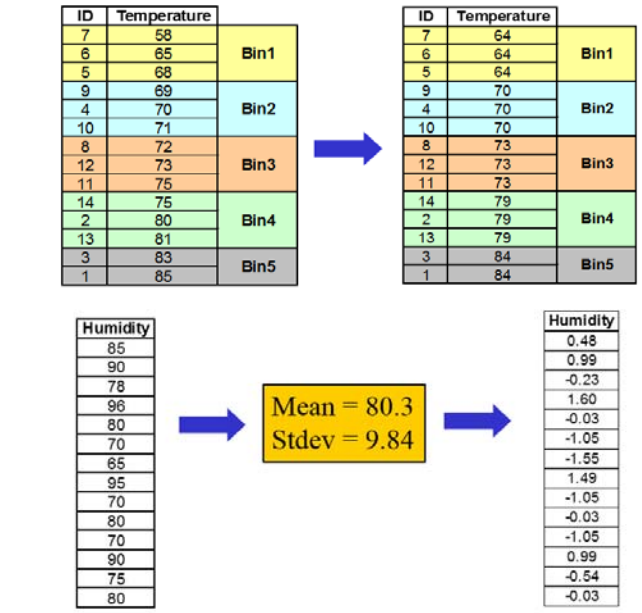

Figure 2: Filling of missing values with binning method: filling of missing values with mean and standard deviation

IV. RESULTS

The results of the binning operations are shown in [Table 1] where Temperature and humidity variables are shown. in T-Bins which stands for Temperature bins show that if the temperature is missing than the sum of three recent values of temperature are to be are to divided by three and the result would be replaced at missing place.

Table 3: Confusion Matrix

	Temperature	Humidity	Wind	Due Point
Temperature	**610**	**17**	**6**	**31**
Humidity	**13**	**720**	**17**	**6**
Wind	**33**	**21**	**522**	**18**
Due Point	**15**	**2**	**4**	**431**
Over all accuracy				**92.57%**

If the average value is replaced than it shows 42.6 (as described in T- Mean Values) degree temperature instead of 41-degree temperature which may become cause of confusion. Same likely the humidity missing variables are to replace as H-Bins quantities. In [Table 2], in case of missing values persisting in wind variables may be replaced with W-Bins and the missing quantities of Dew point are to be replaced with D-Bin quantities. In-order to approximate the accuracy of the proposed system confusion matrix is used [Table 3]. Which shows that 610 observations of temperature are classified and 720 humidity records are classified whereas 522 observations are classified by the classifier and 431 records are classified for the classifier. The measured accuracy of the system is estimated as 92.57%. In [Table 4] comparison with literature have been shown.

Table 1: Binning Operations for filling the missing values by three partitions

Temperature	T-Bins	T-Mean values	Humidity	H-Bins	H-Mean values
40	41	42.6	20	20.3	19.4
41	41	42.6	20	30.3	19.4
42	41	42.6	21	20.3	19.4
42	42.6	42.6	21	20.3	19.4
43	42.6	42.6	20	20.3	19.4
43	42.6	42.6	20	20.3	19.4
44	43.3	42.6	19	18.6.	19.4
44	43.3	42.6	19	18.6	19.4
42	43.3	42.6	18	18.6	19.4
41	42	42.6	20	19.6	19.4
42	42	42.6	21	19.6	19.4
43	42	42.6	18	19.6	19.4

44	44	42.6	18	18	19.4
44	44	42.6	18	18	19.4
44	44	42.6	18	18	19.4

temperature, humidity, wind, wind directions, dew point and other features were found with noise and filled with binning technique. The system is very useful for weather forecasting stakeholders to replace the missing values and to reduce the noise from weather data.

Table 2: Binning Operations for filling the missing values by three partitions

Wind	W-Bins	W-Mean values	Due Point	D-Bins	D-Mean values
11	10	9	0	0	1.6
10	10	9	0	0	1.6
9	10	9	0	0	1.6
9	8.6	9	1	2	1.6
8	8.6	9	2	2	1.6
9	8.6	9	3	2	1.6
7	7.3	9	1	1.3	1.6
7	7.3	9	2	1.3	1.6
8	7.3	9	1	1.3	1.6
13	12	9	3	4	1.6
12	12	9	4	4	1.6
11	12	9	5	4	1.6
10	10	9	1	1	1.6
11	10	9	1	1	1.6
9	10	9	1	1	1.6

Table 4: Status of proposed approach with literature

Approach	Weather (Tem, hum, wind & others)	Thematic map (Yes / No)	Satellite / sensory data pre-processing methods	Machine Learning techniques	Missing Values filled (Yes / No)
1	Smoke, sulfur dioxide	Yes	Primary, Secondary Data	Linear regress (NOAA)	No
2	Tem, hum, wind	No	Weather	ANNs	No
3	Tem, wind, hum, rainfall	Yes	N. A	N. A	No
4	Tem, hum,	No	Sensor data	ANFIS, ANN	No
5	Irrigate	Yes	Sensor data	PLSR	No
6	Tem	No	Primary data		No
7	Tem, hum, wind, Precipitation	Yes	Weather	least-squares regression	No
8	Tem, hum	No	N A	ANNs	No
9	Tem	Yes	N. A	ANNs	No
Our approach					**Yes**

V. CONCLUSION AND DISCUSSION:

A novel weather prediction system has been developed which specifically deals with the noisy data generated by heterogynous and malfunctioning sensors of weather stations. Data cleaning is really important in weather prediction since it provides more precise results. For example; used dataset downloaded from UCI Machine Learning repository was found with missing values where

This research contributes (a) a weather forecasting system, (b) method for filling the missing values and (c) highest accuracy of proposed system This research contributes a weather forecasting system, technique for filling the missing values in weather dataset and highest approximated accuracy of proposed system such as 92.57%.

References

[1] V. B. R. D. and S. Y. Anjaneyulu Yerramilli, "Air Pollution, Modeling and GIS based Decision Support Systems for Air Quality Risk Assessment," Intech open, vol. 2, p. 64, 2018.

[2] M. Matouq et al., "The climate change implication on Jordan: A case study using GIS and Artificial Neural Networks for weather forecasting," J. Taibah Univ. Sci., vol. 7, no. 2, pp. 44–55, 2013.

[3] D. Tien Bui, Q. T. Bui, Q. P. Nguyen, B. Pradhan, H. Nampak, and P. T. Trinh, "A hybrid artificial intelligence approach using GIS-based neural-fuzzy inference system and particle swarm optimization for forest fire susceptibility modeling at a tropical area," Agric. For. Meteorol., vol. 233, pp. 32–44, 2017.

[4] C. Bhargava, V. Kumar Banga, and Y. Singh, "Failure prediction of humidity sensor DHT11 using various environmental testing techniques," J. Mater. Environ. Sci, vol. 9, no. 7, pp. 2009–2016, 2018.

[5] A. G. Mohapatra and S. K. Lenka, "Hybrid decision support system using PLSR-fuzzy model for GSM-based site-specific irrigation notification and control in precision agriculture," Int. J. Intell. Syst. Technol. Appl., vol. 15, no. 1, p. 4, 2016.

[6] L. R. Iverson and A. M. Prasad, "A GIS-Derived Integrated Moisture Index," Charact. Mix. For. Ecosyst. South. Ohio Prior to Reintroduction Fire, pp. 29–41, 1996.

[7] S. Hsu, A. Mavrogianni, and I. Hamilton, "Comparing Spatial Interpolation Techniques of Local Urban Temperature for Heat-related Health Risk Estimation in a Subtropical City," Procedia Eng., vol. 198, no. September 2016, pp. 354–365, 2017.

[8] M. Al-shawwa, A. A. Al-absi, S. A. Hassanein, K. A. Baraka, and S. S. Abu-naser, "Predicting Temperature and Humidity in the Surrounding Environment Using Artificial Neural Network," vol. 2, no. 9, pp. 1–6, 2018.

[9] E. SENER, O. TERZI, S. SENER, and R. KUCUKKARA, "Modeling of Water Temperature Based on GIS and ANN Techniques: Case Study of Lake Egirdir (Turkey)," Ekoloji, vol. 21, no. 83, pp. 44–52, 2012.

[10] https://archive.ics.uci.edu/ml/datasets/SML2010 Accessed on 1-1-2019.