International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320-088X

IMPACT FACTOR: 7.056

IJCSMC, Vol. 11, Issue. 1, January 2022, pg.14 – 24

A Distinctive Approach on the Usage of Edge Computing Concept on Humidity Dataset through Regression Analysis

M. Bhanu Sridhar¹; Y. Srinivas²

¹GVP College of Engineering for Women, Visakhapatnam, AP, India

²GITAM University, Visakhapatnam, AP, India

DOI: 10.47760/ijcsmc.2022.v11i01.004

Abstract:

Edge computing is a paradigm that can distribute the complexity of predictive analysis into smaller pieces, physically placed at the source of contextual information. This allows in processing large amounts of data where it is intricate to use a centralized cloud. Edge Computing makes this possible by taking control of data and services from central hubs, which reduces computational latency on servers. Humidity is one of the main factors that maintain the life of the surface. This article explains how to perform computational analysis at the "edge" by using humidity data sets, and also shows that the most modern data is sufficient for data analysis. Linear Regression and Random Forest Regression algorithms are utilized for data analysis. In addition, this article illustrates the importance of data series for predicting humidity by comparing the analysis of unshuffled data and shuffled data. Metrics have been used to assess the accuracy and point out the importance of sequential data feeds for analysis.

Keywords: Edge Computing, Cloud of Things, Latency, Linear Regression and Random Forest Regression.

1. Introduction:

In today's digital world, every task is performed using data available on the Internet. The data generated from IoT devices is very large and should be transmitted [1] to a data server/cloud, where these large amounts of data can only be analysed on the cloud [2], [3]. The transmission of a large amount of raw data through the network increases the burden on network resources. Edge computing can be regarded as a form of cloud [4], [5], and is the same as a collection or network of micro data centres or repositories. The micro data centres or repositories will store key data and transmit on-demand or on-demand generation of total data. In some cases, this can provide accurate results and consume less network bandwidth [6].

The Internet of Things (IoT) is the most common use case for edge computing. The Internet of Things is all about using edge sensors to collect data from geographically dispersed areas. All "edges" are connected using different networks such as WAN technology. Usually use cellular, VPN and other technologies. The traditional Internet of Things architecture will collect data from sensors scattered around the world and store it in a central repository for storage and analysis [7] [8]. With the increase in the number of devices and hundreds of billions of devices connected, the centralized agent system will inevitably turn into a bottleneck. In order to develop such a system, a large investment is required. On the other hand, the maintenance workload of cloud servers is huge. If the server is unavailable, all subordinate systems will be shut down. Considering these risks and problems, the future of the Internet of Things network largely depends on the decentralization of the network [9].

Humidity refers to the amount of water vapour present in the air. Humidity indicates the presence of precipitation, dew or fog in a certain area. Humidity is an important factor in surface life. Since animal life depends on sweating, sweating is the process of maintaining or regulating body temperature. The higher the humidity, the lower will be the efficiency of heat exchanges. The reason for this is that as humidity increases, it reduces the rate at which moisture evaporates from the skin surface [10]. Humidity can be considered as one of the important factors in predicting the degree of urban pollution. The agricultural sector can also benefit from humidity forecasts.

This article proposes the idea of using humidity data sets to simulate edge computing. The method focuses on how the latest data and a small amount of data are sufficient to predict humidity.

This article also determines that in data sets such as humidity, the data sequence is one of the main factors for predicting the target. By shuffling and ungrouping the entire data set, the same calculation can be performed on the data block. These indicators clearly indicate how the data sequence fed into the algorithm will affect the accuracy of the model.

2. About the Dataset:

The dataset consists of data picked from a monitor system which is placed in a domestic house approximately for 40 days [11]. The data were collected every minute, calculating and uploading it with means.

2.1. Attribute Information:

1. Date in UTC format.
2. Time in UTC format.
3. ComedorSensorForTemperature - Indoor temperature (°C).
4. HabitacionSensorForTemperature - Indoor temperature (°C).
5. TemperatureOfWeather – The weather forecast temperature (°C).
6. ComedorSensorForCO2 - amount of Carbon dioxide in ppm (in dining room).
7. HabitacionSensorForCO2 - amount of Carbon dioxide in ppm (room).

8. ComedorSensorForHumedad- percentage of Relative humidity (dining room).
9. HabitationSensorForHumedad-percentage of Relative humidity (room).
10. ComedorSensorForLighting-lighting (dining room), in Lux.
11. HabitationSensorForLighting-lighting (room), in Lux.
12. Precipitacion - the proportion of the last 15 minutes where rain was detected (in between the range [0, 1]).
13. MeteoExteriorCrepusculo - Sun dusk.
14. MeteoExteriorVientoWind (m/s).
15. MeteoExteriorSolOest – amount of Sunlight in west facade (Lux).
16. MeteoExteriorSolEst – amount of Sunlight in east facade (Lux).
17. MeteoExteriorSolSud - amount of Sunlight in south facade (Lux).
18. MeteoExteriorPiranometro – amount of Sun irradiance (W/m2).
19. EntalpicExterior-1 - state ofenthalpic motor 1, 0 or 1 (on-off).
20. EntalpicExterior-2 - state ofenthalpic motor 2, 0 or 1 (on-off).
21. EntalpicExterior-turbo - state ofEnthalpic motor turbo, 0 or 1 (on-off).
22. ExteriorSensorForTemperature- the amount of Outdoor temperature (°C).
23. ExteriorSensorForHumedad- the percentage of Outdoor relative humidity.
24. DayOfTheWeek has been computed from the date, 1=Monday, 7=Sunday.

3. Pre-Processing on the Dataset:

The data set provided in the repository is in .txt format, which is converted to .csv format for convenience. The data set has 2764 rows, two of which are object types, and the rest are floats.

Fig. I: Details of the Dataset

As a part of pre-processing the data columns 'Precipitation', 'MeteoExteriorSolOest', 'MeteoExteriorSolEst', 'MeteoExteriorSolSud', 'EntalpicExterior-1', 'EntalpicExterior-2', 'EntalpicExterior-turbo ', 'MeteoExteriorViento' are dropped as they have more missing values. The columns 'Date' and 'Time' are also ignored as they have little less significance (dates vary from 13-03-12 to 11-04-2012 and each record is taken per 15 minutes each). The missing values in the columns '5:Weather_Temperature' and 'MeteoExteriorPiranometro' are filled with their respective mean values. After pre-processing, the data has been fragmented into training and testing datasets in the ratio 7:3 for analysis. The analysis methods are explained in the following sections.

4. Regression:

Regression is a technique for describing the statistical connection between two or more variables where any variation is reliant on the variable it is associated with. It also relies on a deviation in one or more independent variables [12].

4.1. Linear Regression:

Linear Regression is a linear model, and commonly used type of predictive analysis in machine learning. Linear Model describes the connection between dependent variable and one or more independent variable(s) [13], [14].

A linear regression can be expressed as

$$Y = c + bX.$$

Here, X is the explanatory variable and Y is the dependent variable. The regression coefficient is b , and c is constant.

The statistics literature refers the above quoted model can have Simple or Multiple model based on the input variables used in the model [15].

Linear regression has many practical applications and they depend on examining two things:

1. Does predictor variables are a good fit in Linear Regression model for predicting the score of outcome variable (dependent variable).
2. If yes, in what way do they indicated by the magnitude and sign of beta estimates and which variables in particular are significant in the prediction of dependent variable [16].

4.1.1. Using Linear Regression on the Dataset:

After using the Linear Regression on the complete dataset the following results were obtained:

Mean Absolute Error: 3.958230558482696

Mean Squared Error: 27.21973339420713

Root Mean Squared Error: 5.217253433963807

Fig. II: Application of Linear Regression on Dataset

4.2. Random Forest Regression:

There are numerous methods for predictive analytics and Random Forest algorithm is one of the most successful methodologies in Machine Learning [17]. It is a type of additive model. Random Forest is a collection of base models. Predictions are made by combining the decisions of base models. [18], [19], [20].

We can represent the Random Forest model in the following way:

$$g(x)=f_0(x)+f_1(x)+f_2(x)+...$$

Here, the ultimate model g is the summation of simple base models f_i . This technique broadly uses multiple models to obtain better performance, this is called model ensembling. All decision trees / base models are constructed independently using a different subsample of the data.

4.2.1. Random Forest Creation pseudo code:

1. Arbitrarily select “k” features from total “m” features, where $k \ll m$
2. Among the “k” features, calculate node “d” by using the best split position.
3. Divide the node into child nodes using the best division methodology.
4. Replicate the above steps till “l” number of nodes has been achieved.
5. Build forest by iterating steps 1 to 4 for “n” times to create “n” trees [22].

4.2.2. Random Forest Prediction pseudo code:

1. Takes in the test data and uses the rules of each randomly formed decision tree to envisage the outcome and accumulates the projected outcome (target).
2. Compute the votes for each predicted target.
3. Consider the highly preferred target as the ultimate prediction from this algorithm.

4.2.3. Using Random Forest Regression on the dataset:

After using the Random forest regression on the complete dataset we get the following results:

Mean Absolute Error: 1.1982855421686744

Mean Squared Error: 3.725649483534699

Root Mean Squared Error: 1.9301941569527918

Fig. III: Application of Random Forest Regression on Dataset

5. Analysis and Results for Simulation of Edge Computing on the Dataset:

There are 2764 rows in the dataset, if we make the dataset folds of 100 rows each we will get 27 sets. The pre-processed data is made to split into 27 sets in two ways, (a) No shuffling of the data provided, (b) Shuffling of data provided.

Both Linear regression and Random Forest Regression are used on both the types of data. The results for un-shuffled data are as shown:

5.1. Linear Regression on Un-shuffled data:

Graphs plotted between test values and predictions for some of the data folds in 27 folds are shown below.

Fig. IV: Application of Linear Regression on Un-shuffled Data

5.2. Random Forest Regression on Un-shuffled data:

Graphs plotted between test values and predictions for some of the data folds in 27 folds are shown below.

Fig. V: Application of Random Forest Regression on Un-shuffled Data

Fig VI: Combined Representation of Algorithms on Unshuffled Data

	MAE	MSE	RMS
0	0.804971	1.173259	1.083171
1	1.240683	2.452159	1.565937
2	0.463783	0.369478	0.607847
3	0.507279	0.484624	0.696149
4	0.909875	1.638583	1.280071
5	1.571740	4.203113	2.050150
6	1.603034	3.612599	1.900684
7	0.987240	1.802504	1.342573
8	0.673732	0.643236	0.802020
9	0.870616	1.110471	1.053789
10	0.826830	1.162487	1.078187
11	0.669870	0.781573	0.884066
12	1.071854	1.649123	1.284182
13	0.375457	0.211004	0.459351
14	1.211300	2.180588	1.476681
15	1.062592	1.473987	1.214079
16	0.487480	0.407700	0.638514
17	0.905675	1.234970	1.111292
18	1.097407	1.752670	1.323885
19	0.546398	0.497197	0.705122
20	0.785583	0.918033	0.958141
21	0.577409	0.470077	0.685622
22	0.571498	0.628325	0.792669
23	0.732335	1.139036	1.067256
24	0.637492	0.693117	0.832536
25	0.955935	1.288578	1.135155
26	0.872646	1.654336	1.286210

Fig. VII: Metrics on un-shuffled data using Linear Regression

	MAE	MSE	RMS
0	0.804971	1.173259	1.083171
1	1.240683	2.452159	1.565937
2	0.463783	0.369478	0.607847
3	0.507279	0.484624	0.696149
4	0.909875	1.638583	1.280071
5	1.571740	4.203113	2.050150
6	1.603034	3.612599	1.900684
7	0.987240	1.802504	1.342573
8	0.673732	0.643236	0.802020
9	0.870616	1.110471	1.053789
10	0.826830	1.162487	1.078187
11	0.669870	0.781573	0.884066
12	1.071854	1.649123	1.284182
13	0.375457	0.211004	0.459351
14	1.211300	2.180588	1.476681
15	1.062592	1.473987	1.214079
16	0.487480	0.407700	0.638514
17	0.905675	1.234970	1.111292
18	1.097407	1.752670	1.323885
19	0.546398	0.497197	0.705122
20	0.785583	0.918033	0.958141
21	0.577409	0.470077	0.685622
22	0.571498	0.628325	0.792669
23	0.732335	1.139036	1.067256
24	0.637492	0.693117	0.832536
25	0.955935	1.288578	1.135155
26	0.872646	1.654336	1.286210

Fig. VIII: Metrics on un-shuffled data using Random Forest Regression

5.3 Linear Regression on Shuffled Data:

Graphs plotted between test values and predictions for some of the data folds in 27 folds are shown below.

Fig. IX: Application of Linear Regression on Shuffled Data

5.4 Random Forest Regression on Shuffled Data:

Graphs plotted between test values and predictions for some of the data folds in 27 folds are shown below.

Fig. X: Application of Random Forest Regression on Shuffled Data

Fig XI: Combined Representation Application of Algorithms on Shuffled Data

	MAE	MSE	RMS		MAE	MSE	RMS
0	3.708664	23.258789	4.822737	0	7.732280	192.698811	13.881600
1	5.663858	73.644283	8.581625	1	6.681550	99.422743	9.971095
2	3.981694	25.029535	5.002953	2	4.083631	23.605345	4.858533
3	4.349412	33.665482	5.802196	3	8.087982	282.428617	16.805613
4	4.778243	34.591312	5.881438	4	5.995128	91.326229	9.556476
5	4.078380	26.974527	5.193701	5	4.487576	30.171993	5.492904
6	5.159859	54.976379	7.414606	6	6.391553	66.098300	8.130086
7	5.396673	46.640841	6.829410	7	6.014345	53.780263	7.333503
8	4.575563	29.308409	5.413724	8	5.161807	43.055653	6.561681
9	3.574396	17.894132	4.230146	9	3.900845	28.001302	5.291626
10	6.376755	88.204743	9.391738	10	6.868704	143.528261	11.980328
11	5.551326	62.532043	7.907720	11	4.981584	43.522030	6.597123
12	5.022867	38.281966	6.187242	12	5.700080	52.947953	7.276534
13	3.553708	22.780469	4.772889	13	5.781628	119.719246	10.941629
14	3.745459	22.977197	4.793454	14	4.294789	28.842425	5.370514
15	6.569922	122.704459	11.077204	15	6.462901	168.191564	12.968869
16	4.996059	45.179499	6.721570	16	5.480791	66.930674	8.181117
17	3.266227	20.894593	4.571060	17	4.336270	33.807110	5.814388
18	3.851404	24.484822	4.948214	18	6.005697	51.697475	7.190096
19	3.382078	19.071901	4.367139	19	3.814673	24.192698	4.918607
20	4.845228	76.474314	8.744959	20	5.254395	50.357096	7.096273
21	4.061977	27.949761	5.286753	21	8.983529	702.943706	26.513086
22	3.583009	25.838481	5.083157	22	10.413554	472.855032	21.745230
23	5.249017	38.821111	6.230659	23	5.426330	36.772950	6.064070
24	4.363831	28.797858	5.366364	24	4.534481	32.006101	5.657393
25	3.811004	22.150361	4.706417	25	4.720078	48.550282	6.967803
26	4.598277	40.544944	6.367491	26	4.598277	40.544944	6.367491

Fig XII: Metrics on shuffled data using Linear Regression

Fig XIII: Metrics on shuffled data using Random Forest Regression

6. Conclusion:

When comparing the results of linear regression and random forest regression, it is clear that random forest regression will provide better results. For regression and classification tasks, Random Forest is a robust and easy-to-interpret machine learning algorithm. One of the main benefits of using random forest is that you can easily check which features or variables contribute to classification or regression, and determine their relative importance based on their depth in the tree. By averaging several trees, Random Forest Technology greatly reduces the risk of overfitting. However, when the data is very small and shuffled, random

forest regression shows poor performance. This is because it requires considerable data to perform complex calculations.

After comparing the analysis results of shuffled and unshuffled data, it is clear that the latest and sequential (i.e., unshuffled) data can provide better results. It can be further deduced that there is little up-to-date data sufficient to understand humidity. Therefore, it can be emphasized that analysing a small amount of the latest data on the edge can reduce the load on the server and provide us with reasonable analysis and reduce latency. The novel concept of using the above algorithms to use edge computing in data collection, temporary storage and operation can be used in different fields of data science and cloud computing, thereby bringing fruitful results to companies and society. It is probable that this kind of idea will present a new opportunity for the upcoming product, and it will be beneficial to the companies and users.

References

- [1]. Chun-Wei Tsai, Chin-Feng Lai, et al., Data Mining for Internet of Things: A Survey, IEEE COMMUNICATIONS SURVEYS & TUTORIALS, VOL. 16, NO. 1, FIRST QUARTER 2014.
- [2]. Ibrahim Abaker, Targio Hashem, and Ibrar Yaqoob, The rise of “big data” on cloud computing: Review and open research issues, Information Systems 47 (2015), Pages 98-115, Elsevier Science Ltd.
- [3]. Jianli Pan, James McElhannon, Future Edge Cloud and Edge Computing for Internet of Things Applications, IEEE Internet of Things Journal.
- [4]. Liang Tong, Yong Li and Wei Gao, A Hierarchical Edge Cloud Architecture for Mobile Computing, IEEE INFOCOM 2016 - The 35th Annual IEEE International Conference on Computer Communications.
- [5]. Quoted from https://en.wikipedia.org/wiki/Edge_computing
- [6]. Quoted from <https://www.networkworld.com/article/3224893/what-is-edge-computing-and-how-it-s-changing-the-network.html>
- [7]. Tessema Mengistu, Abdulrahman Alahmadi, et al., A "No Data Center" Solution to Cloud Computing, 2017 IEEE 10th International Conference on Cloud Computing.
- [8]. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, “Internet of things (IoT): A vision, architectural elements, and future directions,” Future Gen. Comput. Syst., vol. 29, no. 7, pp. 1645–1660, 2013.
- [9]. Quoted from <https://www.bbvaopenmind.com/en/technology/digital-world/3-major-challenges-facing-iot/>
- [10]. Quoted from <https://en.wikipedia.org/wiki/Humidity>
- [11]. Humidity datasets from Machine Learning Repository, <https://archive.ics.uci.edu/ml/datasets.html>
- [12]. Hyung Jun Cho, Seung-Mo Hong, et al., Median Regression Tree for Analysis of Censored Survival Data, IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans (Volume: 38 , Issue: 3 , May 2008), Page(s):715 - 726, Print ISSN: 1083-4427, Electronic ISSN: 1558-2426.
- [13]. H. Shakouri, G. R. Nadimi, et al., Fuzzy linear regression models with absolute errors and optimum uncertainty, 2007 IEEE International Conference on Industrial Engineering and Engineering Management, Print ISBN: 978-1-4244-1528-1, CD-ROM ISBN: 978-1-4244-1529-8 ,Print ISSN: 2157-3611, Electronic ISSN: 2157-362X
- [14]. Wolfgang Tysiak, et al., Regression Analysis of Intrinsic Linear Models with Automated Transformations of Monotone Predictors, 2005 IEEE Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, CD-ROM ISBN: 0-7803-9446-1, Print ISBN: 0-7803-9445-3.
- [15]. Quoted from <https://machinelearningmastery.com/linear-regression-for-machine-learning>
- [16]. Quoted from https://en.wikipedia.org/wiki/Linear_regression
- [17]. Linkai Luo, Weihang Lv ,et al., An empirical study on balance of trade in China based on random forest regression model, 2010 5th International Conference on Computer Science & Education, Electronic ISBN: 978-1-4244-6005-2, Print ISBN: 978-1-4244-6002-1, CD-ROM ISBN: 978-1-4244-6004-5.
- [18]. Quoted from https://turi.com/learn/userguide/supervised-learning/random_forest_regression.htm
- [19]. Deepak Jain, Aradhana Goutam, et al., Optimization of resource and task scheduling in cloud using random forest, 2017 International Conference on Advances in Computing, Communication and Control (ICAC3), Electronic ISBN: 978-1-5386-3852-1, Print on Demand (PoD) ISBN: 978-1-5386-3853-8.
- [20]. Quoted from <http://dataaspirant.com/2017/05/22/random-forest-algorithm-machine-learning>