AN ENSEMBLE LEARNING FRAMEWORK FOR VISIBILITY PREDICTION IN INDO-GANGETIC REGION

Arkapal Panda
Computer Vision and Pattern Recognition Unit
Indian Statistical Institute Kolkata, India
arkapalpanda88in@gmail.com

Vaibhav Kumar [*]**& Tanmay Basu**
Department of Data Science and Engineering,
Indian Institute of Science Education & Research
Bhopal, India
{tanmay,vaibhav}@iiserb.ac.in

ABSTRACT

Visibility of an area can affect all forms of transportation and hence it is important to accurately estimate the visibility of an area for the upcoming days based on different parameters of the meteorological data to take precautions. Several machine learning techniques have been already applied on different kinds of data sets to estimate the visibility, however, none of them were explored on the Indo-Gangetic plane, which witnesses widespread fog primarily during winter. In this spirit, a regression framework is developed to estimate the visibility of the Indo-Gangetic region using the meteorological data, which outperforms the state of the arts.

1 INTRODUCTION

Visibility conditions depend mainly on different phenomena such as changes in air-mass, winds, temperature profiles etc. These phenomena in turn depend on the type of synoptic system affecting a region. Poor visibility condition has an ill effect on the air traffic and transportation system Fabbian et al. (2007), which can badly impact the economy Holtz & Wachs (2011). Moreover, poor visibility due to hazy weather decreases road safety and increases the risk of traffic congestion Gao et al. (2020). There are a lot of research works regarding the estimation of visibility applying machine learning algorithms (mostly regression techniques) using the meteorological data of various regions Gultepe et al. (2006); Li et al. (2017). Ortega et al. (2019) compared five different machine learning classifiers to determine the visibility in the region of Florida in USA. Castillo. B et al. (2022) explored the performance of different classification and regression techniques to predict low visibility events over a data from Mondonedo weather station at Galicia in Spain. There are many other research works for visibility prediction using machine learning and artificial neural network techniques Bari & Ouagabi (2020); Cho & Palvanov (2019). However, to our knowledge, none of these works estimate the visibility of the Indo-Gangetic plane using machine learning, which witnesses widespread fog (a term used for conditions when visibility is less than 1 km), primarily during winter. Despite the large spatial extents, the localised physical nature and spatio-temporal variations of visibility at various scales pose huge challenges in its accurate estimation. Therefore, it is crucial to analyze the factors and their impact on visibility using better data-driven approaches.

2 PROPOSED FRAMEWORK

A regression framework has been developed here to estimate the visibility in the Indo-Gangetic plane in two stages. Note that the meteorological data generally contain mixture of numerical and categorical features and for many features certain values may not be recorded due to system failure. In the first stage, some standard feature engineering schemes has been used to identify potential spatio-temporal features of the meteorological data. The surface meteorological data that is used in this work contain the data from different weather stations in the Indo-Gangetic region, collected from hourly observations contained in US National Oceanic and Atmospheric Administration (NOAA) surface data repository Rutledge et al. (2006). We have grouped the data of each weather station with an unique id to combine them to diminish the effect of the missing values as individual features

[*] Corresponding author

Table 1: Experimental Results of Different Methods for Visibility Estimation

Evaluation Measure	AB	DT	ENet	LGBM	LR	LO	RF	RG	RNN	Proposed
MAE	2.35	1.37	3.10	0.30	3.19	3.12	1.03	2.7	2.21	**0.03**
RMSE	7.77	6.07	8.79	0.45	8.80	8.80	4.99	8.10	1.91	**0.06**

of a particular weather station have less variations than the other stations. Note that the data set has lot of missing values of individual features and hence different standard schemes have been used here to deal with these missing values. The median of all the values of a particular feature is used here to replace the missing values for that feature as the median is a representative of the given feature Lin & Tsai (2020). Similarly for categorical features, mode of a feature instead of median is used to replace the missing values for that feature. Subsequently, to apply the regression techniques on this data set, categorical features are transformed to numerical values following one hot encoding scheme Seger (2018), which is an widely used scheme in this regard. XGboost method Luckner et al. (2017), a state of the art regression model is used to derive relation between different features to estimate the target variable, i.e., visibility. XGboost is an efficient and scalable implementation of gradient boosting technique Chen & Guestrin (2016), which is used here to estimate the visibility from the given data as it performed very well in similar applications Pan (2018); Li et al. (2019).

3 Experimental Analysis

The dataset contains spatiotemporal data of 31 different cities in the Indo-Gangetic plane and it is collected from hourly observations contained in NOAA surface data repository Rutledge et al. (2006). The original dataset has 934807 instances from different weather stations and 121 features for each instance, out of which, 104 features were removed as they have more than 50% missing values. For rest of the features, the missing values were replaced by the median or mode of the other values of the features as stated in section 2. The experimental results are shown in Table 1 which are obtained using these 17 features of the test data. The performance of the proposed framework is compared with the state of the arts viz., linear regression (LR) Montgomery et al. (2021), decision tree (DT) Wantuch (2001), random forest (RF) Kim et al. (2021), lasso (LO) Castillo. B et al. (2022); James et al. (2013), ridge (RG) Uyanık et al. (2021), LightGBM (LGBM) Yu et al. (2021), adaptive boosting (AB) Jakhar et al. (2021), elastic NET (ENET) Castillo. B et al. (2022) and RNN based Jonnalagadda & Hashemi (2020) regression techniques in terms of RMSE and MAE Chai & Draxler (2014); Qi et al. (2020). The proposed method is iterated for 1000 times to avoid the effect of randomization. The data is randomly split into 90% as training data and 10% as test data. The training data is used to tune the parameters of individual regression techniques following 10-fold cross validation technique. Subsequently, the best set of parameters is implemented on the test data for each method. Note that a low score of MAE and RMSE indicate better performance of a regression technique than another method. Table 1 shows that the performance of the proposed method is better than the other techniques in terms of MAE and RMSE. It may be noted that both MAE and RMSE of the proposed one is almost 0, which indicate the method is working reasonably well. These results clearly show the effectiveness of the proposed framework to estimate the visibility of an area. Paired t-test Ruxton (2006) is performed on the scores of proposed framework and other methods in Table 1 to test the statistical significance. The p-value threshold is set to 0.005 here i.e., anything below 0.005 rejects the null hypothesis. It has been found that 16 out of 18 cases are statistically significant when the proposed method beats other techniques in Table 1. The results are inconclusive for the rest two cases in Table 1. Thus it can be concluded that the proposed one performs significantly better than other techniques in 88.88% (16/18) cases in Table 1.

4 Conclusions

The error rate of the proposed method is very low, but still there are scopes to improve the performance further by using the meteorological data of various other regions across globe. Note that the framework is evaluated based on the ground truths of the meteorological data of Indo-Gangetic region. Furthermore, no domain expert is involved in this work and no specific domain knowledge is used to build the framework. We believe that the method has the potential to be used for visibility estimation of any area across globe using the meteorological data. In future, the performance of this framework can be explored on the meteorological data of other regions for visibility estimation.

URM Statement

The authors acknowledge that at least one key author of this work meets the URM criteria of ICLR 2023 Tiny Papers Track.

References

Driss Bari and Abdelali Ouagabi. Machine-learning regression applied to diagnose horizontal visibility from mesoscale nwp model forecasts. *SN Applied Sciences*, 2(4):1–13, 2020.

C Castillo. B, D Casillas-Pérez, C Casanova-Mateo, S Ghimire, E Cerro-Prada, PA Gutierrez, RC Deo, and S Salcedo-Sanz. Machine learning regression and classification methods for fog events prediction. *Atmospheric Research*, 272:106157, 2022.

Tianfeng Chai and Roland R Draxler. Root mean square error (rmse) or mean absolute error (mae)?–arguments against avoiding rmse in the literature. *Geoscientific model development*, 7(3):1247–1250, 2014.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

Young Cho and Akmaljon Palvanov. A new machine learning algorithm for weather visibility and food recognition. *J. Robotics Netw. Artif. Life*, 6(1):12–17, 2019.

Dustin Fabbian, Richard De Dear, and Stephen Lellyett. Application of artificial neural network forecasts to predict fog at canberra international airport. *Weather and forecasting*, 22(2):372–381, 2007.

Kun Gao, Huizhao Tu, Lijun Sun, NN Sze, Ziqi Song, and Heng Shi. Impacts of reduced visibility under hazy weather condition on collision risk and car-following behavior: Implications for traffic control and management. *International journal of sustainable transportation*, 14(8):635–642, 2020.

Ismail Gultepe, Mathias David Müller, and Zafer Boybeyi. A new visibility parameterization for warm-fog applications in numerical weather prediction models. *Journal of applied meteorology and climatology*, 45(11):1469–1480, 2006.

Douglas Holtz and Martin Wachs. Strengthening connections between transportation investments and economic growth. 2011.

Yogesh Kumar Jakhar, Nidhi Mishra, and Rakesh Poonia. Weather event prediction using combination of data mining algorithms. In *Advances in Information Communication Technology and Computing*, pp. 319–326. Springer, 2021.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

Jahnavi Jonnalagadda and Mahdi Hashemi. Forecasting atmospheric visibility using auto regressive recurrent neural network. In *IEEE International Conference on Information Reuse and Integration for Data Science (IRI)*, pp. 209–215. IEEE, 2020.

Bu-Yo Kim, Joo Wan Cha, Ki-Ho Chang, and Chulkyu Lee. Visibility prediction over south korea based on random forest. *Atmosphere*, 12(5):552, 2021.

Shengyan Li, Hong Fu, and Wai-Lun Lo. Meteorological visibility evaluation on webcam weather image using deep learning features. *Int. J. Comput. Theory Eng*, 9(6):455–461, 2017.

Wei Li, Yanbin Yin, Xiongwen Quan, and Han Zhang. Gene expression value prediction based on xgboost algorithm. *Frontiers in genetics*, 10:1077, 2019.

Wei-Chao Lin and Chih-Fong Tsai. Missing value imputation: a review and analysis of the literature (2006–2017). *Artificial Intelligence Review*, 53(2):1487–1509, 2020.

Marcin Luckner, Bartosz Topolski, and Magdalena Mazurek. Application of xgboost algorithm in fingerprinting localisation task. In *IFIP International Conference on Computer Information Systems and Industrial Management*, pp. 661–671. Springer, 2017.

Douglas C Montgomery, Elizabeth A Peck, and G Geoffrey Vining. *Introduction to linear regression analysis*. John Wiley & Sons, 2021.

Luz Ortega, Luis Daniel Otero, and Carlos Otero. Application of machine learning algorithms for visibility classification. In *2019 IEEE International Systems Conference (SysCon)*, pp. 1–5. IEEE, 2019.

Bingyue Pan. Application of xgboost algorithm in hourly pm2. 5 concentration prediction. In *IOP conference series: earth and environmental science*, volume 113, pp. 012127. IOP publishing, 2018.

Jun Qi, Jun Du, Sabato Marco Siniscalchi, Xiaoli Ma, and Chin-Hui Lee. On mean absolute error for deep neural network based vector-to-vector regression. *IEEE Signal Processing Letters*, 27: 1485–1489, 2020.

Glenn K Rutledge, Jordan Alpert, and Wesley Ebisuzaki. Nomads: A climate and weather model archive at the national oceanic and atmospheric administration. *Bulletin of the American Meteorological Society*, 87(3):327–342, 2006. URL `https://www.ncei.noaa.gov/access/search/data-search/global-hourly`.

Graeme D Ruxton. The unequal variance t-test is an underused alternative to student's t-test and the mann–whitney u test. *Behavioral Ecology*, 17(4):688–690, 2006.

Cedric Seger. An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing, 2018.

Tayfun Uyanık, Çağlar Karatuğ, and Yasin Arslanoğlu. Machine learning based visibility estimation to ensure safer navigation in strait of istanbul. *Applied Ocean Research*, 112:102693, 2021.

Ferenc Wantuch. Visibility and fog forecasting based on decision tree method. *Idojárás*, 105:29–38, 2001.

Zhongqi Yu, Yuanhao Qu, Yunxin Wang, Jinghui Ma, and Yu Cao. Application of machine-learning-based fusion model in visibility forecast: A case study of shanghai, china. *Remote Sensing*, 13 (11):2096, 2021.