onglin wei

022300013

Honglin wei

#2022300013

Honglin wei 2024 Submission #2022300013. CONFIDENTIAL REVIEW COPY. DO NOT DISTRIBUTE.

Predicting Road Visibility in Shenzhen Using Machine Learning

Anonymous Honglin wei submission

Paper ID 2022300013

Abstract

This study focuses on predicting road visibility in Shen-zhen by leveraging real-time image data from the city’sroad monitoring system alongside meteorological data us-ing machine learning methodologies. Initially, historicaldata were analyzed to identify correlations between fac-tors such as precipitation, humidity, and wind speed withroad visibility, with distribution maps illustrating these re-lationships. Subsequently, various machine learning algo-rithms, including Support Vector Regression (SVR) and De-cision Tree Regression, were employed to construct predic-tive models. The performance of these models was eval-uated using metrics such as Mean Squared Error (MSE),with results indicating that SVR provides better predictiveaccuracy compared to Decision Tree Regression.

This research offers an effective approach for predictingroad visibility in Shenzhen and provides valuable insightsfor meteorological monitoring and traffic management insimilar urban settings. The code available at https://github.com/zhuiber/MachineLearning/

tree/main/Term%20paper

Index Terms Road visibility prediction, machine learn-ing, SVR, Decision Tree, urban traffic, Shenzhen.

1. Introduction

In modern urban traffic management, the prediction andmonitoring of road visibility are of paramount importance.Visibility, as a common indicator in daily life, plays a cru-cial role in driving and transportation. Especially in rapidlydeveloping cities like Shenzhen, with high traffic densityand a large number of vehicles, timely and accurate predic-tion of road visibility is particularly crucial.

1.1. Background

In Shenzhen, a thriving modern city in China, manag-ing traffic safely amid rapid growth is crucial. Visibility onroads directly affects people’s safety and daily lives.

Machine learning offers new ways to predict road visibil-ity. By using weather data and machine learning techniques,

Figure 1. Haze weather in Shenzhen. Retrieved from Internet

we can accurately predict visibility, giving early warningsfor accidents and guiding traffic management decisions.

1.2. Existing methods

1. Hybrid Model of CNN and LSTM

Zhang, Ling, Zhao Yang, and Jian Zhang proposed ahybrid model combining Convolutional Neural Networks(CNN) and Long Short-Term Memory (LSTM) networksfor road visibility estimation [6]. The CNN extracts spatialfeatures from images, while the LSTM captures temporaldependencies in the data. This hybrid approach leveragesboth spatial and temporal information, leading to high ac-curacy in visibility prediction. It is particularly robust indynamic and changing environments.

2. Real-time Visibility Estimation Using Deep Neural

Networks

He, Yong, Wei Zhang, Lei Xu, and Xin Li developed adeep neural network model specifically designed for real-time visibility estimation in foggy conditions [2]. Themodel is trained on a large dataset of foggy weather imagesand demonstrates high accuracy and fast processing speed,making it suitable for real-time applications. The model ef-fectively handles varying levels of fog density.

3. Visibility Estimation of Expressways Based on Deep

Learning

Qi, Chao, Jingwen Zhang, Mingjie Huang, and Bo Wangintroduced a deep learning model trained on expressway

onglin wei

022300013

Honglin wei

#2022300013

Honglin wei 2024 Submission #2022300013. CONFIDENTIAL REVIEW COPY. DO NOT DISTRIBUTE.

images to estimate visibility [4]. The model architecturefocuses on capturing the features specific to highway envi-ronments. It achieves high accuracy in visibility estimationon expressways, demonstrating the potential of deep learn-ing in specialized transportation contexts.

4. Multi-scale Convolutional Neural Networks for Real-

time Visibility Estimation

Shi, Qianqian, Yanfeng Han, and Xing Liu proposed amulti-scale CNN approach, where features are extracted atdifferent scales to capture both fine and coarse details inthe images [5]. This multi-scale approach enhances themodel’s ability to predict visibility accurately across vari-ous weather conditions and environments, providing robustreal-time performance.

5. Foggy Road Visibility Estimation Using Generative

Adversarial Networks (GANs)

Liu, Yiming, Qian Sun, and Jian Wang employed Gener-ative Adversarial Networks (GANs) to improve the visibil-ity of road images by generating de-fogged images, whichare then used for visibility estimation [3]. The GAN-basedmethod shows significant improvements in image clarityand visibility estimation accuracy. It excels in extremelyfoggy conditions where traditional methods may struggle.

6. Conclusion

However, common limitations include high computa-tional demands, data specificity, and potential generaliza-tion issues. Future research should focus on developingmore generalized models that can handle diverse weatherconditions and road types while optimizing computationalefficiency.

2. Machine learning methodology

In my study, I employ traditional machine learning al-gorithms (Support vector regression, Decision tree). Thisselection allows for simpler implementation and less com-putational demand compared to deep learning methods.

2.1. Support Vector Regression

Support Vector Regression (SVR) is a type of regres-sion technique that uses Support Vector Machines (SVM) tofunction as a regression estimator. The main principle be-hind SVR is to identify a hyperplane in a high-dimensionalspace that has the maximum margin with respect to thetraining data. In other words, SVR aims to find the hy-perplane that minimizes the generalization error while stillfitting the training data well.

The equation for SVR can be represented as:

y = wTx+ b

where:- y is the predicted output value - w is the weight

vector - x is the input sample - b is the bias term

Figure 2. One-dimensional linear SVR . Adapted from [1]

The goal of SVR is to find the optimal values of w andb that minimize the error between the predicted output andthe true output while also maximizing the margin betweenthe hyperplane and the training data.

Additionally, SVR uses a kernel function to map the in-put data into a higher-dimensional space in order to find ahyperplane that can separate the data points. The most com-monly used kernel functions in SVR are the linear kernel,polynomial kernel, and radial basis function (RBF) kernel.

The optimization problem for SVR can be formulated as:

min

w,b,ξ,ξ∗

||w||2+ C

N∑

i=1

(ξi +ξ∗i )

subject to the constraints:

yi − wTxi − b ≤ϵ+ξi,wTxi + b− yi ≤ϵ+ξ∗i ,

ξi,ξ

∗

i ≥0,

where:- N is the number of training samples -ξi,ξ

∗

are slack variables that allow for some deviation from themargin - C is the regularization parameter -ϵ is the margin

of tolerance - w is the weight vector - b is the bias term

In summary, the principle of SVR involves finding theoptimal hyperplane that minimizes the error while maxi-mizing the margin with respect to the training data. Theuse of kernel functions allows SVR to handle non-linear re-lationships between the input data and the output, making ita powerful tool for regression tasks.

2.2. Decision tree Regression

A 1D regression with decision tree.

The decision trees is used to fit a sine curve with additionnoisy observation. As a result, it learns local linear regres-sions approximating the sine curve.

We can see that if the maximum depth of the tree (con-trolled by the max-depth parameter) is set too high, the de-cision trees learn too fine details of the training data andlearn from the noise, i.e. they overfit.

onglin wei

022300013

Honglin wei

#2022300013

Honglin wei 2024 Submission #2022300013. CONFIDENTIAL REVIEW COPY. DO NOT DISTRIBUTE.

Figure 3. Decision tree.Retrieved from https://scikit-

learn.org/

3. Experiments

3.1. Data Collection, Preprocessing and Analysis

Data Collection from the website:

https :// opendata . sz . gov . cn / data /

dataSet/toDataDetails/29200\_00903518

This dataset contains hourly telemetry data from Shen-zhen, with 3,730 records and 64 fields. The data types areprimarily integers and strings. Some example fields includewind direction, cloud height, relative humidity, datetime,surface minimum temperature, grassland maximum tem-perature, automatic precipitation amount, minimum stationpressure, maximum wind speed, and more.

Collection Time: The timestamps in the dataset rangefrom August 9,2015, to April 6,2020, depending on thespecific record.

1. We find that some characteristic sets have a lot of miss-ing values, so we first delete these sets.

2. Since the processed dataset still has some missing val-ues, we choose to replace the missing values with 0.

3. We calculate the correlation values between visibilityand other characteristic sets.

4. We observe that the dataset has many characteristics,and many of them have low correlation with visibility.Therefore, we only select characteristics with a corre-lation value greater than 0.15.

5. We explore the distribution for every characteristic.

6. We calculate the mean and variance for each charac-teristic. The results are shown in Table 2.

Feature Correlation

RELHUMIDITY 0.311094

MINRELHUMIDITY 0.306294

INSTANTWINDV 0.212944

HEXMAXWINDV 0.198315

WINDV10MS 0.162007

MAXWINDV10MS 0.156455

AUTOPRECIPAMOUNT 0.156363

GRASSLANDMAXTEMP 0.152857

Table 1. Correlation of various features with VISIBILITY.

Figure 4. Correlation with Visibility

Figure 5. VISIBILITY Figure 6. RELHUMIDITY

Figure 7. MINRELHUMID-

ITY Figure 8. INSTANTWINDV

onglin wei

022300013

Honglin wei

#2022300013

Honglin wei 2024 Submission #2022300013. CONFIDENTIAL REVIEW COPY. DO NOT DISTRIBUTE.

Figure 9. HEXMAXWINDV Figure 10. WINDV10MS

Figure 11.

MAXWINDV10MS

Figure 12. GRASSLAND-

MAXTEMP

Table 2. Mean and Standard Deviation

Feature Mean Standard Deviation

VISIBILITY 220.496110.7851

RELHUMIDITY 72.9619316.2497

MINRELHUMIDITY 70.9512116.75499

INSTANTWINDV 24.7168915.29268

HEXMAXWINDV 49.7613921.20455

WINDV10MS 19.829499.918758

MAXWINDV10MS 27.1217211.91181

GRASSLANDMAXTEMP 266.4373109.4651

Figure 13. SVR Regression with PCA

3.2. Support Vector Regression

To develop an SVR (Support Vector Regression) algo-rithm to predict visibility using the given features, we will

follow these steps:

1. Load the dataset from the provided Excel file.2. Extract one principal component using PCA.

Figure 14. Decision Tree Regression with PCA

The SVR model predictions form a smoother curvethrough the data, indicating that the model is captur-ing a general trend rather than fitting the noise. Thered line follows a trend with less variance compared tothe Decision Tree model, suggesting a more general-ized model.

3. Split the data into training and testing sets.4. Train the SVR model using the training set.

5. Evaluate the model using 10-fold cross-validation on

the training set and compute the average mean squarederror (MSE).

6. Print the learned model and the average MSE.

3.3. Decision tree aggression

To develop an Decision tree aggression algorithm to pre-dict visibility using the given features, we will follow these

steps:

1. Load the dataset from the provided Excel file.2. Extract one principal component using PCA.3. Split the data into training and testing sets.

4. Train the Decision Tree model using the training set.5. Evaluate the model using 10-fold cross-validation on

the training set and compute the average mean squarederror (MSE).

6. Print the learned model and the average MSE.

The Decision Tree model’s predictions create a dense,crisscrossed web of red lines connecting the actual visibil-ity values with the predicted values. This suggests that theDecision Tree model has a high variance, fitting the trainingdata very closely (possibly overfitting). The model capturesa lot of noise from the training data.

onglin wei

022300013

Honglin wei

#2022300013

Honglin wei 2024 Submission #2022300013. CONFIDENTIAL REVIEW COPY. DO NOT DISTRIBUTE.

Figure 15. SVR Predictions vs Actual Values on Testing Set

4. Results and Analysis

4.1. Support Vector Regression

Metric MSE

Training Mean Squared Error (MSE)11006

Testing Mean Squared Error (MSE)15326

Table 3. Mean Squared Error (MSE) for Training and Testing in

svr

The MSE is higher on the testing set compared to thetraining set, indicating that the SVR model performs bet-ter on the data it was trained on and shows some degreeof overfitting. The difference between training and testingMSE suggests that while the model generalizes reasonablywell, there is still a noticeable performance drop on unseendata.

Trend Line: The SVR predictions (red dots) show asmoother, more continuous trend line, which indicates thatthe SVR model captures a general trend or pattern in thedata.

Error Distribution: The predicted values are generallycloser to the actual values (green dots), especially in themiddle range of the principal component values. The spreadof the predictions appears narrower compared to the Deci-sion Tree model.

Outliers: There are fewer extreme outliers in the SVRpredictions compared to the Decision Tree model. The SVRmodel seems to handle outliers better, keeping most predic-tions within a reasonable range from the actual values.

4.2. Decision tree aggression

Similar to the SVR model, the Decision Tree modelshows a higher MSE on the testing set compared to thetraining set, indicating overfitting. However, the MSE val-ues are higher overall compared to the SVR model, suggest-

Metric MSE

Training Mean Squared Error (MSE)21697

Testing Mean Squared Error (MSE)23787

Table 4. Mean Squared Error (MSE) Results in decision tree

Figure 16. Decision Tree Predictions vs Actual Values on Testing

Set

ing that the Decision Tree model may not be as effective atcapturing the underlying patterns in the data.

Trend Line: The Decision Tree predictions (red dots)do not form a clear trend line. Instead, they appear morescattered around the actual values.

Error Distribution: The predicted values are spread outmore widely around the actual values. This suggests thatthe Decision Tree model may be overfitting, capturing morenoise in the training data which leads to more variability inpredictions.

Outliers: There are more extreme outliers and a widerspread of predictions compared to the SVR model. Thisindicates that the Decision Tree model is less robust to noiseand outliers in the data.

4.3. Comparison of Two Methods

Overall, the SVR model demonstrates better perfor-mance in terms of MSE for both training and testing setscompared to the Decision Tree model. While both mod-els exhibit overfitting, the SVR model’s lower MSE valuesindicate it is more effective at modeling the data. However,the larger generalization gap in the SVR model suggests thatfurther tuning or regularization might be needed to improveits generalization capability.

The SVR model captures the underlying trend moresmoothly, has a narrower error distribution, and handlesoutliers more effectively. The Decision Tree model, on theother hand, shows signs of overfitting with more scatteredpredictions and a higher number of extreme outliers.

onglin wei

022300013

Honglin wei

#2022300013

Honglin wei 2024 Submission #2022300013. CONFIDENTIAL REVIEW COPY. DO NOT DISTRIBUTE.

4.4. Comparison with existing methods

Decision Tree Regression with PCA and SVR with PCAare simpler and more interpretable models but tend to havehigher error rates and may overfit or underfit depending onthe data complexity.

Deep Learning Approaches (e.g., Hybrid CNN-LSTM,multi-scale CNNs, GANs) demonstrate superior perfor-mance in capturing complex patterns, handling variousweather conditions, and providing robust real-time perfor-mance. These methods, however, require more computa-tional resources and larger datasets for training.

In scenarios where computational resources and trainingdata are abundant, deep learning models significantly out-perform traditional methods like Decision Trees and SVRin terms of accuracy and robustness, especially in dynamicand challenging environments.

By comparing the results, it is evident that deep learn-ing models offer substantial improvements in visibility esti-mation tasks over traditional machine learning approaches,albeit at the cost of increased complexity and resource re-quirements.

5. Conclusion

In this study, we explored the prediction of road visibilityin Shenzhen using machine learning methods. We collectedand analyzed meteorological data to identify key factors af-fecting visibility, such as humidity and wind speed. Twomachine learning algorithms, Support Vector Regression(SVR) and Decision Tree Regression, were implemented tocreate predictive models.

Our results showed that SVR outperformed DecisionTree Regression in terms of predictive accuracy, with lowerMean Squared Error (MSE) on both training and testingdatasets. While both models exhibited some degree of over-fitting, the SVR model captured the underlying trends moreeffectively and demonstrated better generalization to newdata.

Overall, this research highlights the potential of tradi-tional machine learning techniques for predicting road visi-bility, providing valuable insights for urban traffic manage-ment and meteorological monitoring. Future work couldexplore more advanced models and larger datasets to fur-ther improve predictive performance.

References

[1] M. Awad, R. Khanna, M. Awad, et al. Support vector regres-

sion. In Efficient Learning Machines: Theories, Concepts,and Applications for Engineers and System Designers, pages

67-80. Apress, New York, NY, 2015. 2

[2] Yong He, Wei Zhang, Lei Xu, and Xin Li. Real-time visibil-ity estimation in foggy weather using deep neural networks.

IEEE Access, 8:36212-36223, 2020. 1

[3] Yiming Liu, Qian Sun, and Jian Wang. A novel approach for

foggy road visibility estimation using generative adversarial

networks. Applied Soft Computing, 107:107379, 2022. 2

[4] Chao Qi, Jingwen Zhang, Mingjie Huang, and Bo Wang. Visi-

bility estimation of expressways based on deep learning. Jour-nal of Transportation Safety Security, 13(2):191-210, 2021.[5] Qianqian Shi, Yanfeng Han, and Xing Liu. Multi-scale con-volutional neural networks for real-time visibility estimation.

Neural Computing and Applications, 33:4371-4382, 2021. 2[6] Ling Zhang, Zhao Yang, and Jian Zhang. Road visibility esti-

mation using a hybrid model of convolutional neural networksand long short-term memory. Journal of Intelligent Trans-

portation Systems, 25(4):345-358, 2021. 1