Çukurova Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi, 35(3), ss. 567-573, Eylül 2020
Çukurova University Journal of the Faculty of Engineering and Architecture, 35(3), pp. 567-573, September 2020

Performance Evaluation of Various Regression Models and Features for Prediction of Ozone Concentration

Sezer DÜMEN[1], Ercan AVŞAR[*1], Ulus ÇEVİK[1]

[1]*Çukurova University, Faculty of Engineering, Department of Electrical and Electronics Engineering, Adana*

Geliş tarihi: 24.07.2020 *Kabul tarihi: 23.10.2020*

Abstract

Air pollution caused by ozone is a problem which threaten human health. Therefore, prediction of O_3 concentration is important. In this work, O_3 concentration level for Adana, Turkey is predicted with support vector regression (SVR), multi-layer perceptron (MLP), gradient boosting decision trees (GBDT), K nearest neighbors (KNN), elastic net machine learning methods. Parameters utilized for this prediction are hourly measurement of pollutants like particular matter (PM10), sulfur dioxide (SO_2), nitrogen dioxide (NO_2), nitrogen oxides (NO_x), nitric oxide (NO) concentrations and also meteorological parameters like air temperature, wind speed, relative humidity, air pressure, wind direction. Additionally, hour, day and season information are used as features. It has been shown that SVR method achieves the best result with R^2 value of 0.9697. Furthermore, backward elimination method is implemented for feature selection process and according to the results, current O_3 concentration has the highest importance to predict the concentration for the next hour.

Keywords: Air quality, Ozone concentration, Machine learning, Regression

Ozon Konsantrasyonu Tahmininde Çeşitli Regresyon Modelleri ve Özniteliklerin Performans Değerlendirmesi

Öz

Ozondan kaynaklanan hava kirliliği insan sağlığını tehdit eden bir problemdir. Bu nedenle, O_3 konsantrasyonunun tahmini önemlidir. Bu çalışmada, Türkiye'nin Adana ili için O_3 konsantrasyon seviyesi, destek vektör regresyonu (DVR), çok katmanlı algılayıcı (ÇKA), gradyan artırılmış karar ağaçları (GAKA), K en yakın komşu (KEK) ve elastik net makinesi öğrenme yöntemleri kullanılarak tahmin edilmiştir. Bu tahmin için kullanılan parametreler, partiküler madde (PM10), sülfür dioksit (SO_2), azot dioksit (NO_2), azot oksitler (NO_x), azot monoksit (NO) gibi kirleticilerin konsantrasyonları ve ayrıca hava sıcaklığı, rüzgâr hızı, bağıl nem, hava basıncı, rüzgâr yönü gibi meteorolojik parametrelerin saatlik ölçümleridir. Ek olarak saat, gün ve sezon bilgileri de parametre olarak kullanılmaktadır. DVR yöntemi ile elde edilen R^2 değeri 0,9697 olup diğer yöntemlerle elde edilen değerlerden yüksektir. Ayrıca öznitelik seçimi için geriye doğru eleme yöntemi uygulanmıştır ve sonuçlara göre bir sonraki saatin O_3 konsantrasyonunu tahmin etmek için şimdiki O_3 konsantrasyon seviyesinin en önemli öznitelik olduğu görülmüştür.

Anahtar Kelimeler: Hava kalitesi, Ozon konsantrasyonu, Makine öğrenmesi, Regresyon

[*]Sorumlu yazar (Corresponding author): Ercan AVŞAR, *ercanavsar@cu.edu.tr*

1. INTRODUCTION

Tropospheric ozone (O_3) is an air pollutant that has negative effects on human health and climate change [1,2]. Since O_3 is not released directly into the air, it is considered as a secondary pollutant. With industrialization without emissions controls, the release of precursors to O_3, such as nitrogen oxides (NO_x) and volatile organic compounds (VOCs), are gradually increasing [3]. In addition, O_3 formation is related with some meteorological parameters like temperature, wind speed, wind direction and rainfall. [4].

Negative effects of O_3 and the other air pollutants are admitted worldwide. Thus, limits for hazardous levels of these air pollutants are officially determined by the governments. In Turkey, these limits are set as 500 µg/m^3 for sulfur dioxide (SO_2), 400 µg/m^3 for nitrogen dioxide (NO_2), 240 µg/m^3 for O_3. [5]. Besides, having high concentration of these gases in the air is dangerous for those people having health conditions like asthma or chronic obstructive pulmonary disease (COPD). Considering this information, determination of O_3 concentration in the atmosphere is an important problem not only for air pollution analysis, but also for medical purposes.

Various machine learning methods have been utilized in the literature for estimation of pollutant gas concentrations. For instance, in some of the previous works, threshold values or concentration ranges have been determined for air pollutants and air pollution estimation have been done with classification methods such as decision trees, random forest classification, and artificial neural networks (ANN) [6-8]. In addition, various regression methods were used for concentration estimates [9-13]. For instance, root mean square error (RMSE) values obtained through support vector regression (SVR), ANN and decision tree methods were compared in estimation of O_3, SO_2 and NO_2 concentration levels [9]. It was shown that SVR produces the lowest RMSE value. In another study, regression models such as linear regression, ridge regression, multilayer perceptron (MLP), Elman neural network were used for prediction of O_3 level [10]. Among these methods, the best RMSE result was obtained with the MLP method. Besides, the results of a study handled in Canada shown that extreme learning machine (ELM) method outperformed ANN and multiple linear regression (MLR) to predict O_3, particular matter 2.5 micrometers or less in diameter (PM2.5) and NO_2 concentration levels [11]. In another comparative study, ANN was shown to be giving better results than MLR method in predicting the maximum hourly O_3 concentration for the next day [12]. On the other hand, algorithms utilizing deep learning methods are mentioned in more recent studies. For example, a new deep learning model to forecasting of O_3 concentration was proposed by researchers in Aarhus, Denmark [13]. According to the results of the study, this new method outperformed SVR, ANN and MLR methods.

Instead of predicting the parameters related with air pollution, it is also possible to analyze the correlation between the pollutant gases. For example, relationship of O_3 concentration with other gas concentrations and meteorological parameters are analyzed using ANN and support vector machines (SVM) methods [14]. It has been verified that the O_3 concentration was negatively correlated with carbon monoxide (CO), nitric oxide (NO) and NO_x as these pollutants are known precursors of O_3.

In this study, O_3 concentration for Adana, Turkey is estimated using five different machine learning methods. These methods are SVR, MLP, gradient boosting decision tree (GBDT), K nearest neighbors (KNN), and elastic net. Measurements for various gases (particular matter 10 micrometers or less in diameter (PM10), SO_2, NO_2, NO_x, NO, O_3) and meteorological parameters together with temporal data are used as features. Additionally, backward feature elimination method is used for determining the most appropriate predictors for O_3 estimation and select a subset of features that are more useful in modeling O_3 level.

2. MATERIALS AND METHODS

2.1. Machine Learning Methods

SVM is a supervised learning method used for binary classification problems [15]. In this method, a given dataset is non-linearly mapped to a higher dimensional feature space. For classification problems, in this feature space, data is separated by hyperplanes with the maximum margin between class boundary. SVM can also be used for regression (SVR) or multi-label classification problems.

The performance of an SVR model highly depends on choice of kernel function for mapping and parameters C and ε. In this work Gaussian kernel function is used. This kernel has only one parameter, γ, which is associated with the width of the Gaussian function. On the other hand, C and ε parameters control the tradeoff between the variance and bias in the training process. Therefore, careful selection of these parameters is important. For this purpose, best combination of these parameters is searched in a grid space. As a result, optimum parameters are found as 100 for C, 0.01 for γ and 0.1 for ε.

MLP is a feed-forward ANN [16]. MLP network consists of the input, hidden, and output layers with their nodes called neurons [17]. There can be multiple hidden layers between the input layer and the output layer. The inputs of each neuron are multiplied by a value called weight and all of them are summed up and passed through the activation function. Widely used activation functions are sigmoid, step, linear, hyperbolic tangent and rectified linear unit (ReLu). The output of this activation function is transferred to the next layer in the same way. Each node in one layer is connected to each node in the next layer. Also, back-propagation method is used for updating the weights and reduce the error. In this method, number of neurons in hidden layer can affect the performance of model. The network structure used in this work consist of one hidden layer containing 50 neurons.

Elastic net is a regularization and variable selection method. Elastic net regression is mixture of lasso and ridge regression models. This method is particularly useful when the number of features is much bigger than the number of data samples and when there are correlated features [18]. The formulation of the elastic net regression model is as where λ_1 and λ_2 are regularization parameters, X is sample matrix, y is response vector, β is weight vector (Equation 1).

$$\hat{\beta} = \operatorname{argmin}_{\beta} |y - X\beta|^2 + \lambda_2 |\beta|^2 + \lambda_1 |\beta| \quad (1)$$

For this method, parameter selection is applied to choose the best λ_1 and λ_2 parameters and optimum values of these parameters are found as 0,009 for λ_1 and 0,0005 for λ_2.

GBDT is a machine learning method for regression and classification problems [19]. Considering decision trees as weak learners, GBDT algorithm is a strong learner created by combining many decision trees. A single decision tree tends to overfitting, while a GBDT reduces the risk of overfitting [20]. Each new decision tree in the gradient boosting algorithm is created using the residuals from the previous step.

Number of weak learners is the key parameter that affects the performance of the GBDT method. This parameter should be well adjusted in order to prevent overfitting. After trying several values, it has been determined that 200 is appropriate for this parameter.

KNN algorithm is another machine learning method used in this study. Due to its simplicity and straightforward idea, it is a very common method [21]. In this method, the distance values between a test sample and all training samples are calculated. The K nearest samples to the test sample is used for classification or regression. The key part in this algorithm is to choose the optimum value of K. For this study, models have been trained with K values between 1 and 5 and optimum K value is found as 4.

2.2. The Dataset

The dataset was obtained from the air monitoring database of the Republic of Turkey Ministry of Environment and Urbanization [22]. The download settings were adjusted to include hourly data for gas concentrations such as PM10, SO_2, NO_2, NO_x, NO, O_3 and meteorological parameters such as air temperature, wind speed, relative humidity, air pressure and wind direction. All data between the years 2016 and 2020 years were downloaded. However, the proportion of missing data was very high. After removing the samples containing missing data, total of 11 weeks of data was obtained which is used for prediction of O_3 concentration. Hourly change of the gas concentrations for one week for duration of the data is shown in Figure 1.

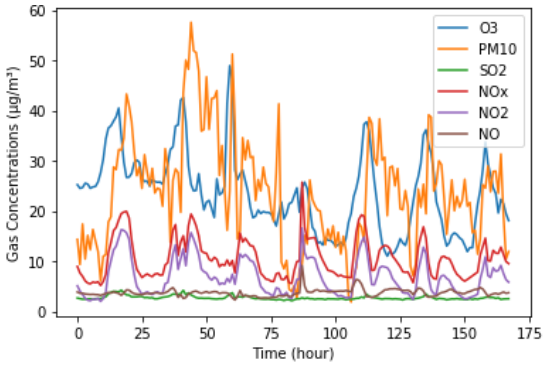

Figure 1. Hourly change of gas concentrations for one week of data

Since the pollutant gas concentrations vary with time, the involvement of temporal features to the dataset may have positive effect on the prediction performance. Therefore, three temporal information (day, hour and season information) were added as new features by using one hot encoding method. In one-hot encoding method, categorical variables are represented as vectors. The related element in the set of categorical variables is represented by 1 and other elements by 0. Hence, in this study, 24 new columns were added to represent "hour of day" information, 7 new columns were added to represent "day of week" information and 4 new columns were added to represent "season of year" information. Using the timestamp of each row, the related hour, day and season column was set as 1 while other columns were assigned 0.

In addition to the temporal features, O_3 concentrations belonging to five different previous measurements were added as new features. These previous data are O_3 values for one week ago, one day ago, three hours ago, two hours ago and one hour ago of the O_3 concentration to be predicted. Eventually, a dataset containing 1848 rows and 51 columns was obtained.

2.3. Experiments

70% of the dataset was randomly selected as training set and the remaining part was left as test set. Since the continuous features in the dataset had been on different scales, these features were scaled through z-score normalization given by

$$Z = \frac{X - \mu}{\sigma} \tag{2}$$

where μ, σ are mean and standard deviation of the feature vector X, respectively (Equation 2).

In order to estimate O_3 concentration for the next hour, SVR, MLP, GBDT, KNN, elastic net models were trained with the optimum parameters given in section 0. In addition, R^2, RMSE, mean absolute error (MAE), mean absolute percentage error (MAPE) metrics were used for performance evaluation of the models.

In addition to estimation of O_3 concentration using all 50 features, a subset of features giving higher test performance was searched through backward feature elimination method. This method is a feature selection process where one feature is removed at every step [23]. The difference in test performance caused by individual removal of every single feature is observed and the feature whose removal generates the higher performance is discarded. These steps are then repeated for the remaining features until the desired number of

features is achieved or until the desired error rate is reached.

3. RESULTS AND DISCUSSION

Prediction performances of the machine learning methods trained using all the features are given in Table 1. As can be seen in Table 1, the highest R^2 value is achieved by the SVR method followed by elastic net and GBDT methods, respectively (Figure 2). Additionally, the lowest RMSE, MAE and MAPE values are obtained by SVR as well. KNN and MLP methods have poor performance compared with the others for the problem in this study. While MLP has better performance than KNN in RMSE and R^2 metrics, KNN has better performance than MLP in MAE and MAPE metrics.

Table 1. Prediction results of the methods when all features are used

	RMSE	**MAE**	**MAPE**	**R^2**
SVR	**4.8119**	**2.9961**	**7.8820**	**0.9697**
GBDT	5.0216	3.2637	8.8934	0.9670
MLP	5.2853	3.6139	9.9880	0.9635
KNN	5.5198	3.5452	9.5762	0.9601
Elastic net	5.0178	3.2144	8.5179	0.9671

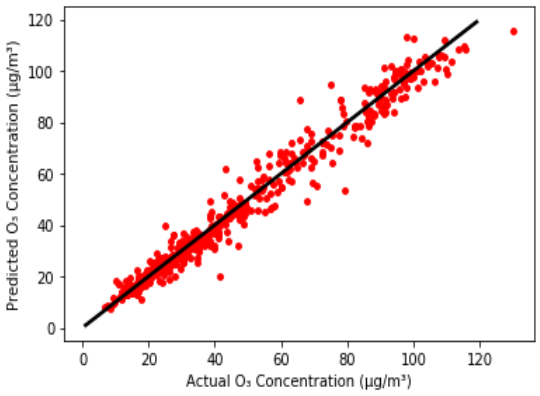

Figure 2. Predictions versus actual O_3 concentrations for SVR method

The R^2 values at every step of backward elimination for all of the methods are given in Table 2. For SVR method, removal of PM10 and relative humidity features does not affect the performance so these two features are considered to be redundant. Besides, removal of a feature does necessarily cause a decrement in the performance. This is true when a feature does not represent the target value well. In other words, presence of such features makes it more complicated to model the data, hence causes low performance. Some examples of this situation can be seen in Table 2. For instance, performance of MLP method is increased when PM10 feature is removed. Similarly, for the same feature, higher R^2 values are obtained in GBDT and KNN methods. Furthermore, the same single feature, which is current O_3 concentration, has remained at the final step of the backward elimination in all of the methods. This means that the current O_3 concentration has the highest importance on prediction of the O_3 concentration for the next hour.

4. CONCLUSION

Air pollution has harmful impact on human health and environment. Estimation of air pollutants concentration is a major significance for any country. In this work, different machine learning methods are used for prediction of hourly O_3 concentration in Adana, Turkey. Although, the performance measures of all methods used in this study are satisfactory and their results are close each other, SVR outperformed other machine learning methods.

Furthermore, feature selection is implemented to find the best feature subsets and analyze the effect of features on the performance of models. The results indicate that the current O_3 concentration has the highest importance for predicting the O_3 concentration of the next hour. It has been observed that using only the current O_3 concentration as a single feature can yield R^2 values above 0.95.

Table 2. Feature selection results. *(RH: Relative humidity, AP: Air pressure, WS: Wind speed, WD: Wind direction, T: Temperature, 1wpO_3: One week previous O_3, 1dpO_3: One day previous O_3, 2hpO_3: Two hour previous O_3, 1hpO_3: One hour previous O_3, Date: All temporal features generated via one hot encoding)*

SVR		MLP		GBDT		KNN		Elastic net	
Removed feature	R^2 after removed	Removed feature	R^2 after removed	Removed feature	R^2 after removed	Removed feature	R^2 after removed	Removed feature	R^2 after removed
PM10	0.9697	1wpO_3	0.9690	1dpO_3	0.9678	1hpO_3	0.9604	SO_2	0.9673
RH	0.9697	PM10	0.9705	NO_x	0.9675	PM10	0.9611	1wpO_3	0.9673
AP	0.9695	RH	0.9712	RH	0.9678	SO_2	0.9626	WS	0.9674
SO_2	0.9693	AP	0.9717	WD	0.9681	1dpO_3	0.9625	RH	0.9674
NO	0.9691	NO	0.9708	NO	0.9681	AP	0.9624	NO_x	0.9674
NO_2	0.9690	SO_2	0.9710	AP	0.9674	NO	0.9619	T	0.9674
WS	0.9687	2hpO_3	0.9709	2hpO_3	0.9678	T	0.9589	PM10	0.9673
1wpO_3	0.9686	T	0.9724	1wpO_3	0.9675	NO_2	0.9583	NO	0.9673
NO_x	0.9685	Date	0.9694	PM10	0.9681	RH	0.9563	AP	0.9671
T	0.9683	WD	0.9680	1hpO_3	0.9677	NO_x	0.9515	NO	0.9669
1hpO_3	0.9677	WS	0.9675	WS	0.9658	2hpO_3	0.9477	2hpO_3	0.9667
WD	0.9670	NO_x	0.9677	SO_2	0.9658	WD	0.9406	WD	0.9664
Date	0.9650	NO_2	0.9653	NO_2	0.9650	Date	0.9451	Date	0.9651
1dpO_3	0.9604	1dpO_3	0.9611	T	0.9644	WS	0.9459	1dpO_3	0.9608
2hpO_3	0.9534	1hpO_3	0.9531	Date	0.9448	1wpO_3	0.925	1hpO_3	0.9532

Besides, PM10 and relative humidity are found to least important features as they are removed at initial steps of the backward elimination.

In addition, this work can be extended by using of parameters like solar radiation, rainfall, CO concentration and this may improve the performance of prediction models. Also, collection of more and clean data will definitely allow for deeper analysis.

5. REFERENCES

1. Lippmann, M., 1989. Health Effects of Ozone a Critical Review. Japca, 39(5), 672-695.

2. Manning, W.J., Tiedemann, A.V., 1995. Climate Change: Potential Effects of Increased Atmospheric Carbon Dioxide (CO_2), Ozone (O_3), and Ultraviolet-B (UV-B) Radiation on Plant Diseases. Environmental Pollution, 88(2), 219-245.

3. Selin, N.E., Wu, S., Nam, K.M., Reilly, J.M., Paltsev, S., Prinn, R.G., Webster, M.D., 2009. Global Health and Economic Impacts of Future Ozone Pollution. Environmental Research Letters, 4(4), 044014.

4. Vukovich, F.M., Sherwell, J., 2003. An Examination of the Relationship Between Certain Meteorological Parameters and Surface Ozone Variations in the Baltimore–Washington Corridor. Atmospheric Environment, 37(7), 971-981.

5. Hava Kalitesi Değerlendirme ve Yönetimi Yönetmeliği. Turkish Official Journal (Issue: 29940). https://www.resmigazete.gov.tr/eskiler/2008/06/20080606-6.htm

6. Yu, R., Yang, Y., Yang, L., Han, G., Move, O.A., 2016. RAQ–A Random Forest Approach for Predicting Air Quality in Urban Sensing Systems. Sensors, 16(1), 86.

7. Corani, G., Scanagatta, M., 2016. Air Pollution Prediction Via Multi-label Classification. Environmental Modelling & Software, 80, 259-264.

8. Rybarczyk, Y., Zalakeviciute, R., 2016. Machine Learning Approach to Forecasting Urban Pollution. In 2016 IEEE Ecuador

Technical Chapters Meeting (ETCM) IEEE, 1-6.

9. Shaban, K.B., Kadri, A., Rezk, E., 2016. Urban Air Pollution Monitoring System with Forecasting Models. IEEE Sensors Journal, 16(8), 2598-2606.

10. Salazar-Ruiz, E., Ordieres, J. B., Vergara, E.P., Capuz-Rizo, S.F., 2008. Development and Comparative Analysis of Tropospheric Ozone Prediction Models Using Linear and Artificial Intelligence-based Models in Mexicali, Baja California (Mexico) and Calexico, California (US). Environmental Modelling & Software, 23(8), 1056-1069.

11. Peng, H., Lima, A.R., Teakles, A., Jin, J., Cannon, A.J., Hsieh, W.W., 2017. Evaluating Hourly Air Quality Forecasting in Canada with Nonlinear Updatable Machine Learning Methods. Air Quality, Atmosphere & Health, 10(2), 195-211.

12. Chaloulakou, A., Saisana, M., Spyrellis, N., 2003. Comparative Assessment of Neural Networks and Regression Models for Forecasting Summertime Ozone in Athens. Science of The Total Environment, 313(1-3), 1–13. doi:10.1016/s0048-9697(03)00335-8

13. Ghoneim, O.A., Manjunatha, B.R., 2017. Forecasting of Ozone Concentration in Smart City Using Deep Learning. In 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI) (pp. 1320-1326). IEEE.

14. Luna, A.S., Paredes, M.L.L., De Oliveira, G.C.G., Corrêa, S.M., 2014. Prediction of Ozone Concentration in Tropospheric Levels Using Artificial Neural Networks and Support Vector Machine at Rio de Janeiro, Brazil. Atmospheric Environment, 98, 98-104.

15. Smola, A.J., Schölkopf, B., 2004. A Tutorial on Support Vector Regression. Statistics and Computing, 14(3), 199-222.

16. Gardner, M., Dorling, S., 1998. Artificial Neural Networks (the multilayer perceptron)- a Review of Applications in the Atmospheric Sciences. Atmospheric Environment, 32(14-15), 2627–2636. doi:10.1016/s1352-2310(97)00447-0.

17. Govindaraju, R.S., Rao, A.R. (Eds.). 2013. Artificial Neural Networks in Hydrology (Vol. 36). Springer Science & Business Media.

18. Zou, H., Hastie, T., 2005. Regularization and Variable Selection Via the Elastic Net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2), 301-320.

19. Friedman, J.H., 2001. Greedy Function Approximation: a Gradient Boosting Machine. Annals of Statistics, 1189-1232.

20. Mohan, A., Chen, Z., Weinberger, K., 2011. Web-search Ranking with Initialized Gradient Boosted Regression Trees. In Proceedings of the Learning to Rank Challenge, 77-89.

21. Wu, X., Kumar, V., Quinlan, J.R., Ghosh, J., Yang, Q., Motoda, H., Zhou, Z.H., 2008. Top 10 Algorithms in data Mining. Knowledge and Information Systems, 14(1), 1-37.

22. Republic of Turkey Ministry of Environment and Urbanization National Air Quality Monitoring Network. https://www.havaizleme.gov.tr/

23. Kohavi, R., John, G.H., 1997. Wrappers for Feature Subset Selection. Artificial Intelligence, 97(1-2), 273-324.