Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/50/20241617

Detecting concentration of Ozone based on different machine learning methods

Chihao Yu

University of California, San Diego, 9500 Gilman Dr, La Jolla, CA 92093

chy007@ucsd.edu

Abstract. Hundreds of years ago, the environment was not that polluted; however, as the industrial revolution came, although our life became much easier, many toxic substances were released. These toxic substances result in Ozone pollution. Thus, the topic of the research focuses on predicting the amount of ozone present in the atmosphere in some areas of the world and how different factors, like environmental issues, can affect the pollution of Ozone. Moreover, which factors contribute the most. Through the extensive literature review and retrieval, it raised the question of how different methods of detecting concentrations of Ozone have drawbacks and advantages. In short, the main conclusion is that the simple methods (KNN and Perceptron) we are familiar with are enough to produce good enough results, and though these two simple methods have some disadvantages, there are more advanced methods (GBRT) that make the results even better.

Keywords: KNN (k-nearest neighbors algorithm), GBRT (Gradient Boosted Regression Trees), Perceptron, Ozone, Machine learning

1. Introduction

Right now, as more and more industrial activity becomes more enhanced, the problems with Ozone (O_3) become more serious as it has started to negatively affect the lives of people. With more Ozone Depletion, the UV Radiation has caused more cases of skin cancer. The production of Ozone comes from NOx and VOCs [1]. It is easy for these two reactants to form secondary pollutants under the sunlight. It can easily oxidize the furniture and out everyday objects. Most importantly, it can irritate people's respiratory system and nervous system.causing respiratory and neurological diseases, and ozone can even cause chromosomal diseases, harming people's life. In each body part, the paper introduces one method that is provided by other papers and then the paper introduces the pros and cons of each method. For the first body paragraph, the paper introduces the KNN (k-nearest neighbors algorithm) method to predict the amount of Ozone. In the second paragraph, the paper introduces the Perceptron method to predict the amount of ozone. For the third paragraph, the paper introduces the GBRT (Gradient Boosted Regression Trees) method to predict the amount of Ozone. This not only provides future researchers the chance to revise the current method to increase accuracy, but perhaps leveraging the pros and cons of each method to design a much better solution.

DOI: 10.54254/2755-2721/50/20241617

2. KNN

The first method is about the KNN. The KNN is basically classifying the data points based on how close the new data points are to a specific group of data. In this paper, 2535 valid data is chosen from Dayton, Galveston, and Brazoria Areas, with some impactful factors taken into account. And the factors considered are temperature, wind speed,relative humidity,sea level pressure, and precipitation. For the temperature, it is usually peak temperature and temperature from different pressures. For the wind speed, including storm intensity, mean wind speed, mean wind vector, and individual. East-west, north-south under pressure (different altitudes) and Directional wind speed. For the Relative humidity. Including K-index, average humidity under different pressures. Then it is the sea level pressure, since the temperatures it accompanies directly correlate with the sea level pressure. Lastly, the precipitation since it is related to relative humidity and the humidity can impact the concentration of Ozone. Then there is a chart showing whether given input that is already known and the output, we can classify the new data point using the shortest distance. For example, maybe given some factors, this particular set of data is classified as Ozone Pollution, while others are not. Then we can classify new data based on K nearest data points. And if in these K nearest data points, more are classified as Ozone pollution than no Ozone pollution, then the new data points will be classified as Ozone pollution and vice versa. And we can also standardize the distance to standardize the degree of impact on different sets of factors using z-score. In this study, not reaching pollution is represented 0 while reaching pollution is represented 1. As a result of this study, the KNN is proven to be really effective, it has the accuracy value of 94% [2]. Well, the advantage of K-NN is it does not need the training process and moreover, it is really easy to implement– we only need to calculate the distance between new data points and old data points. The downside of K-NN is that since it needs to compare the new data point with the old classified data point, with large datasets, it may take a long time. And also, the prediction complexity is high, as there may be tens of thousands of dimensions. Moreover, it is sensitive to the outliers [3].

3. Perceptron method

The second method is about the perceptron method, which basically takes in some number of inputs and then gives weights to each input. Then it adds in the bias terms to give some level of flexibility. And then it determines the results using activation functions such as the sigmoid function and the ReLu function. We adjust the weights and biases as more data comes in using some kinds of formulas. In this paper, the author introduces that originally, it uses Gaussian mode calculation and gray prediction. The gray prediction method is to enhance the order of the original sequence through gray generation, so as to get the forecast result of the given original sequence through the reduction operation based on the fitting process of the generated sequence using various models. The Gaussian Method is to calculate using the Space rectangular coordinate system X,Y and Z. However, using gaussian method, this methods need to meet series of ideal conditions and the real application is hard to fulfill these ideal conditions, thus the prediction is not accurate. Researchers also used the gray prediction method to regard the atmospheric environment as a GM (1, 1) model is established for a gray system, and then the model is constructed by weakening the randomness. The method of establishing continuous calculus equations is used to predict environmental pollution. But this method can only make macro predictions, and the prediction values do not have sufficient credibility and other defects.

In this paper, the author chose 2535 lists of data in Galveston and Brazoria County from 1998 to 2004 as the training data, and after getting the exact perceptron formula, it was confirmed that this formula is reliable by test. They put the values of U50, V50, HT50, K1, etc. into the perceptron and clearly get a binary classification. Again, just like the method of K-NN, we use similar sets of features and update the weights as more data comes in and we can get a pretty good set of weights.[4] The downside of the perceptron method is that it can only train linearly separable models, but we know that in real life, the data sets are hardly linearly separable. Moreover, just like KNN, Perceptron is the binary classification method, which can only have two results [5].

DOI: 10.54254/2755-2721/50/20241617

4. GBRT

The third method is the GBRT method. It is basically a non-parametric statistical learning technique for classification and regression. The basic idea is through multiple iterations, we ensemble multiple weak learners to build a strong learner. The basic learner of GBRT is a regression tree. It divides the characteristic space into different parts, assigns values to each part, and then makes the prediction. Different from Boosting, The GBRT is intended to reduce the residual value provided by the previous model and build the basic learner based on that. In this passage, it uses ground-based observation,weather data and other supporting data such as NDVI in MODIS,emission, population,DEM and LUCC. And the way of measuring the data is quite efficient and reliable.For example, when it measures the real O_3 concentration level, it uses the highest O_3 level in terms of 8-hour sliding window. For the weather data, it also uses the average 24 hour data. To better predict the data, they get rid of the feature that has little effect on the data,such as population agitation pressure,elevations and land use type. They use the chart to analyze as more features are gotten rid of, the effect it has on slope, the decisive factor root-mean-square error and intercept. And they found out getting these 4 features can make their data sets better. After that, they split the data into 10 sets. They use one set for validation sets and 9 sets as training sets. They took the average of 9 sets and 1 validation sets into the linear fit regression. Compared to random forest and FTWR, GBRT performs much better than these two [6]. The advantages of GBRT are it can deal with large datasets, it can deal with both numerical and categorical data and it can handle missing data (no imputation is required). Moreover, it can deal with different kinds of data,strong prediction ability and can fully utilize the characteristic information of all pixels in the entire study area. There are also clearly some disadvantages, for example, it can cause overfitting and it is expensive and time-consuming (it requires a lot of trees).

5. Conclusion

This paper selects three other papers and analyzes the methods that were discussed by these papers. The methods chosen by the three papers are KNN, perceptron, and GBRT. And I chose these in a way that is both accessible to some beginners in machine learning and to more advanced readers in the machine learning field. With KNN, it is easy to understand and implement, but the prediction process may be slow since it needs to be compared one by one. For the perceptron, it is again to be understood, but the single-layer perceptron has limited expressive power, and the results are not accurate. For the GBRT method, although it requires advanced study and is possible to result in overfitting, it can deal with different kinds of data and has strong prediction ability. And in this way, we should define much more advanced methods that can leverage these advantages and disadvantages. Certainly, this paper has some disadvantages; only a subset of the methods are evaluated, and there will probably be some other methods that already accomplish. And the author could include more of the latest methods so that the conclusion of effectiveness for future work is that the paper should focus more on analyzing more advanced methods, evaluating their disadvantages, and possibly designing more efficient and effective methods of measuring the concentration of Ozone.

References

[1] Ziyang Zhang (2018) Research on ozone pollution problem based on perceptron model. China Strategic Emerging Industry,10 (053-058).

[2] Qiwen Wang (2019) Research on environmental pollution problems based on machine learning. China High-Tech,1 (118-121).

[3] Yifei Li, Qin Kai, Ding Li, Wenzhi Fan, He Qin, (2020) Ground-level ozone concentration estimation based on gradient boosting regression tree algorithm. China Environmental Science,Vol.40 (3): 997-1007.

[4] Sillman, S. (2003). Overview: Tropospheric ozone, smog and ozone-NOx-VOC sensitivity. Treatise on Geochemistry, 14, 2022.

[5] Networks, N. (2023, March 12). What are the advantages and disadvantages of using a single-la yer perceptron versus a multi-layer perceptron?. Perceptrons: Single-layer vs Multi-layer Neu

DOI: 10.54254/2755-2721/50/20241617

ral Networks. https://www.linkedin.com/advice/0/what-advantages-disadvantages-using-single-layer#:~:text=One%20of%20the%20main%20disadvantages,straight%20line%20to%20the%20data.

[6] Chatterjee, M. (2023, June 5). A quick introduction to KNN algorithm. Great Learning Blog: Free Resources what Matters to shape your Career! https://www.mygreatlearning.com/blog/knn-algorithm-introduction/.