大數據應用與管理
Journal of Big Data Application and Management
2024 Vol. 4

Using Machine Learning to Predict AQI

Yu-Chun Fan [1], Chih-Chung Yang [2*]

ABSTRACT

This paper will delve into the topic of "how various factors in the air affect the Air Quality Index (AQI)". Air quality is closely related to people's daily lives, and the air quality index, as an indicator to quantitatively assess the degree of air pollution, can intuitively reflect the health of the environment. Therefore, this study aims to use machine learning methods to analyze a large amount of historical data in the past, explore and quantify various factors that affect air quality, and then predict future air quality indexes.

Keyword: Air Quality Index, Machine Learning

1. INTRODUCTION

1.1. Introduction of Air Pollution

Air pollution can cause health effects. Short-term health effects include the harm of eyes, nose, and throat, headaches, allergic reactions, and upper respiratory infections. Some long-term health effects, such as lung cancer, brain damage, liver damage, kidney damage, heart disease, and respiratory disease. Unexpectedly, it can even cause diabetes. The effects of air pollution for the environment are haze, eutrophication, and global climate change. The current criteria pollutants are Carbon Monoxide (CO), Lead (Pb), Nitrogen Dioxide (NO_2)), Ozone (O_3)), Particulate matter (PM), Sulfur Dioxide (SO_2)). The Air Quality System (AQS) contains ambient air pollution data collected by EPA, state, local, and tribal air pollution control agencies from thousands of monitors. The methodology is that EPA regulations require state environmental agencies to report air monitoring data at least quarterly. The monitoring agencies must report the measured data, along with metadata about the site, monitoring equipment, and associated quality assurance data to the AQS. Data for a given calendar quarter are due to EPA by the end of the following quarter. Office of Air Quality Planning and Standards (OAQPS): Its primary mission is to preserve and improve air quality in the United States. To accomplish this, OAQPS:

1. Comply and review air pollution data.
2. Develop regulations to limit and reduce air pollution.
3. Assist states and local agencies with monitoring and controlling air pollution.
4. Make information about air pollution available to the public.
5. Report to Congress the status of air pollution and the progress made in reducing it.

Since different pollutants have different effects, the National Ambient Air Quality Standards (NAAQS) are also different. Some pollutants have standards for both long-term and short-term averaging times. The short-term standards are designed to protect against acute or short-term health effects, while the long-term standards were established to protect against chronic health effects. The air quality index (AQI), which quantifies air quality in a region, is used to express how polluted the air is currently or how polluted it is forecasted to become. There are five steps of AQI:

1. When AQI is between 0-50, the air pollution level is excellent. (Green)
2. When AQI is between 51-100, the air pollution level is good. (Yellow)
3. When AQI is between 101-150, the air pollution level is lightly polluted. (Orange)
4. When AQI is between 151-200, the air pollution level is moderately polluted. (Red)
5. When AQI is between 201-300, the air pollution level is heavily polluted. (Purple)
6. When AQI is over 301, the air pollution level is severely polluted. (Dark Red)

1.2. Domestic AQI and International AQI

1.2.1 The International Air Quality Index

The Air Quality Index (AQI) is calculated based on monitoring data for various pollutants in the air on a given day. These pollutants include ozone (O_3), fine particulate matter (PM2.5), particulate matter (PM10), carbon monoxide (CO), sulfur dioxide (SO_2), and nitrogen dioxide (NO_2). Each pollutant is assigned a sub-index value based on its impact on human health. The AQI for a specific monitoring station on a given day is determined by the highest sub-index value among these pollutants. The calculation of AQI among different countries is essentially the same and is based on this formula：

$$IAQI_P = \frac{IAQI_{Hi} - IAQI_{Lo}}{BP_{Hi} - BP_{Lo}}(C_P - BP_{Lo}) + IAQI_{Lo}$$

$IAQI_P$ = the Air Quality Sub-Index for the respective region for pollutant parameter P,
C_P = the pollutant concentration of pollutant parameter P in the respective region,
BP_{Hi} = the concentration breakpoint of pollutant parameter P that is $\leq C_P$,
BP_{Lo} = the concentration breakpoint of pollutant parameter P that is $\geq C_P$,
$IAQI_{Hi}$ = the index breakpoint corresponding to BP_{Hi}

[1]范榆君 國立清華大學 學士生 fannjimmy0811@gmail.com
[2]楊至中 臺灣知識庫 研究員 ccyang.phd@gmail.com

$IAQI_{Lo}$ = the index breakpoint corresponding to BP_{Lo}

1.2.2 Domestic AQI

After the real-time concentration of each measurement item is calculated according to the following formula, the six items including O_3, PM2.5, PM10, CO, SO_2, and NO_2 can be obtained according to the table below. The real-time secondary index value of the measurement item is taken, and the maximum value is taken as the real-time air quality index. The maximum value measurement item is the index pollutant:

1. $O_{3,8h}$: Take the most recent 8-hour moving average.
2. O_3 : Get the instant concentration value.
3. PM2.5 : 0.5 × average of the previous 12 hours + 0.5 × average of the previous 4 hours.
4. PM10 : 0.5 × average of the previous 12 hours + 0.5 × average of the previous 4 hours.
5. CO : Take the most recent 8-hour moving average.
6. SO_2 : Get the instant concentration value.
7. $SO_{2,24h}$: Take the latest continuous 24-hour concentration average.
8. NO_2 : Get the instant concentration value.

Air pollution can cause health effects. Short-term health effects incl

2. LITERATURE REVIEW

2.1. Machine Learning Method

2.1.1 Linear Regression

To predict the most accurate value of y for a given **x**, we need to estimate the relationship between **x** and **y** using known data examples.

In linear regression, the model used to predict **y** is a linear function. Consider the linear function defined as $f_{\mathbf{w},\mathbf{b}}(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + \mathbf{b}$, $\mathbf{w} \in \mathbb{R}^{D}$, $\mathbf{y} \in \mathbb{R}$, where w is a weight vector and **b** is a bias term. To find the optimal $\mathbf{w}^*$ and $\mathbf{b}^*$, we minimize the following loss function, which measures the difference between the predicted and actual values: $\frac{1}{N}\sum_{i=1}^{N}(f_{\mathbf{w},\mathbf{b}}(\mathbf{x}_i) - \boldsymbol{y}_i)^2$. By minimizing this loss function, we can estimate the relationship between **x** and **y** using the linear function.

The reason we use the squared loss function (binary function) instead of the absolute value function is that the squared loss is differentiable, making it easier to find the minimum value by using differentiation techniques. On the other hand, the absolute value function is not differentiable at some points, complicating the optimization process.

Although higher-order functions could potentially fit the data better, they are more sensitive to noise in the data, which might result in predictions that deviate significantly from the actual values.

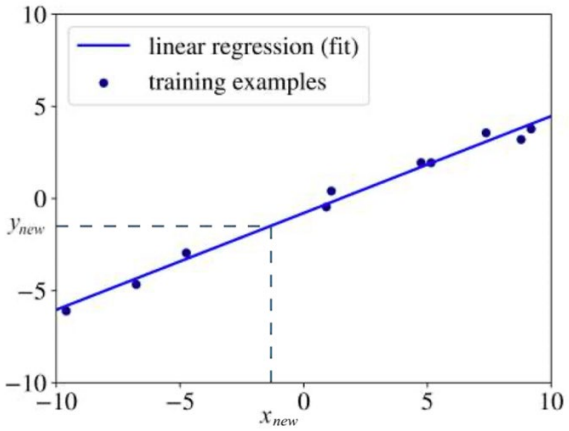

Figure 1: Linear regression of degree 2

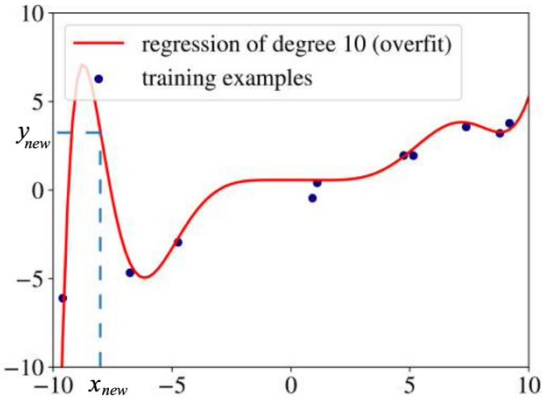

Figure 2: Linear regression of degree 10

2.1.2 Logistic Regression

In logistic regression, for each $i \in \{1,2,\ldots,N\}$, $\boldsymbol{y}_i$ is either 0 or 1. Define y=0 if y is negative, y=1 otherwise.

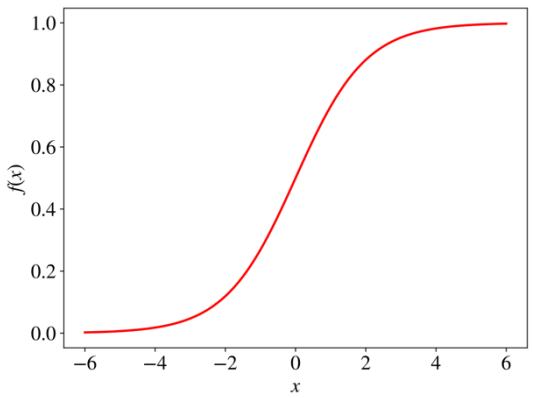

Figure 3: Standard logistic function

2.1.3 Decision Tree

A decision tree is an acyclic graph that can be used to make decision. In every branching node of the graph has a specific value k, which examined $\mathbf{x}_i^{(j)}$ for some $j \in \{1,2,\ldots,D\}$ where $i \in \{1,2,\ldots,N\}$. If $\mathbf{x}_i^{(j)} < k$, then the left branch will be followed, otherwise, the right side will be followed.

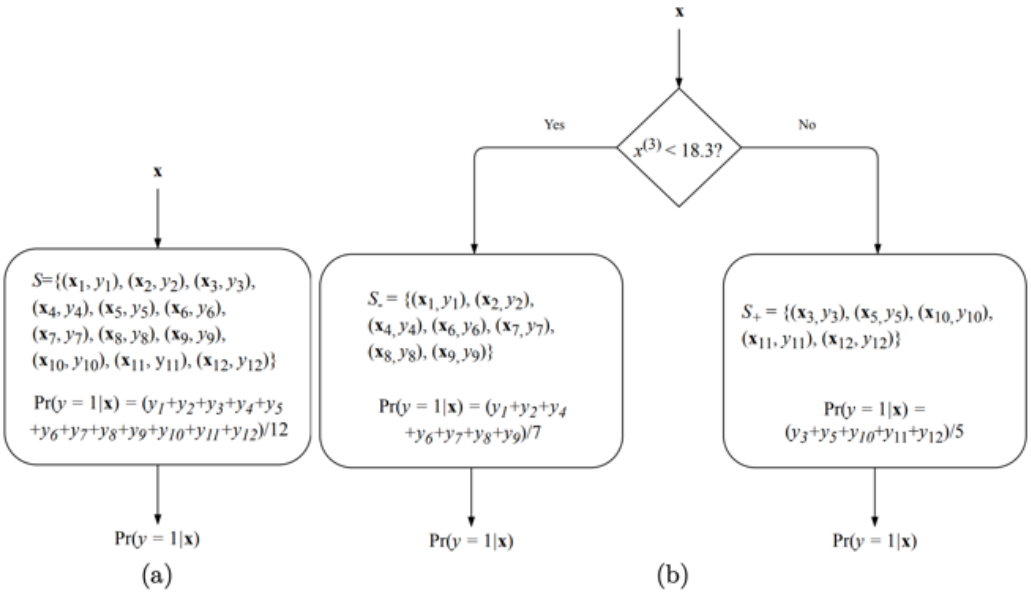

Figure 4: An illustration of a decision tree building algorithm

2.1.4 k-Nearest Neighbors

k-Nearest Neighbors (kNN) unlike other learning algorithm that can discard the training data after the modal was made. kNN keeps all the training data, when a previously unseen example **x** was input, kNN will find k training examples which is closest to **x**, and return majority label or average label to **x**'s label **y**. We define the distance function here by following:

$$d(x_i, x_j) = \frac{\sum_{k=1}^{D} x_i^{(k)} x_j^{(j)}}{\sqrt{\sum_{k=1}^{D} \left(x_i^{(k)}\right)^2} \sqrt{\sum_{k=1}^{D} \left(x_j^{(k)}\right)^2}}$$

3. RESEARCH METHODS

3.1. Collecting dataset

To accurately predict AQI, we first need real AQI data, which serves as our target variable. The data on the left side of the AQI column—such as precipitation, wind speed, wind direction, temperature, relative humidity, and ultraviolet radiation—are factors we believe have a direct impact on AQI. These environmental variables are collected from the Central Weather Administration. The data on the right side of the AQI column consists of various pollutants, as AQI is calculated based on the concentrations of these pollutants. This pollutant data is sourced from the Ministry of Environment.

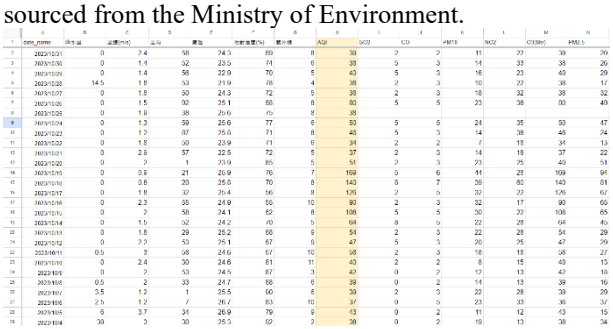

Figure 5: collected data

3.2. Training and Testing

According to the paper: "Empirical studies show that the best results are obtained if 80 percent of the data is used for training." and" Empirical studies show that the best results are obtained if the remaining 20 percent of the data is used for testing. "Consider a dataset of size N, denoted by $\{\mathrm{x_i}, \mathrm{y_i}\}_{\mathrm{i=1}}^{\mathrm{N}}$. For each $i \in \{1,2,\ldots,N\}$, $x_i = \left(x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(D)}\right)$, $x_i^{(j)} \in \mathbb{R}$ is called a feature vector with D-dimension and $y_i \in \mathbb{R}$ is called a real-valued target.

3.3. Support Vector Machine (SVM)

In SVM, for each $i \in \{1,2,\ldots,N\}$, $\mathrm{y_i}$ is either -1 or 1. We want to find a hyperplane h such that if $\mathrm{y_i} \neq \mathrm{y_j}$, then $(\mathbf{x_i}, \mathrm{y_i})$ and $(\mathbf{x_j}, \mathrm{y_j})$ lies on different side of h. The hyperplane h is given by two parameters w and b as following: h:**wx**+b=0 where $\mathbf{w} \in \mathbb{R}^{\boldsymbol{D}}$, $\mathrm{b} \in \mathbb{R}$.

3.3.1 Definition of h

The hyperplane h splits the examples if and only if the hyperplane h determines by $(\mathbf{w}, \mathrm{b})$ satisfies $\min_{i=1,2,\ldots,N} y_i(\boldsymbol{w}\boldsymbol{x_i} + b) = 1$.

3.3.2 Target of SVM

The problem that we want the machine to solve is minimize $\|\boldsymbol{w}\|$ subject to $\min_{i=1,2,\ldots,N} y_i(\boldsymbol{w}\boldsymbol{x_i} + b) = 1$ which is equivalent to minimize $\|\boldsymbol{w}\|$ subject to $\min_{i=1,2,\ldots,N} y_i(\boldsymbol{w}\boldsymbol{x_i} + b) \geq 1$. Which is also called Hard margin SVM.

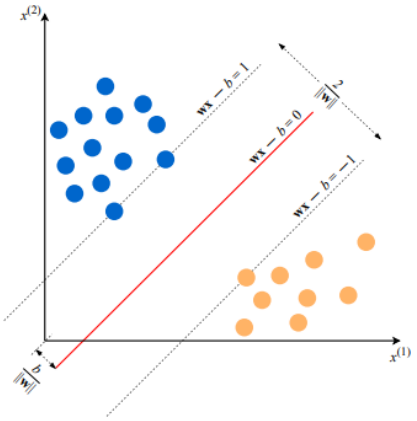

Figure 6: SVM (Hard margin SVM)

3.4. Confusion Matrix

When evaluating the performance of a model, we often use confusion matrix to give an analysis. So, we first define:

1. True positive (TP) = A test result that correctly indicates the presence of an attribute
2. True negative (TN) = A test result that correctly indicates the absence of an attribute
3. False positive (FP) = A test result which wrongly indicates that a particular attribute is present
4. False negative (FN) = A test result which wrongly indicates that a particular attribute is present

	Actual positive	Actual negative
Predicted positive	TP	FP
Predicted negative	FN	TN

Figure 7: Confusion Matrix

In general, we initially use Accuracy to know how well our prediction models are. Where Accuracy is defined by $\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}$.

4. ANALYSIS AND RESULTS

4.1. Using RapidMiner to run the model of SVM

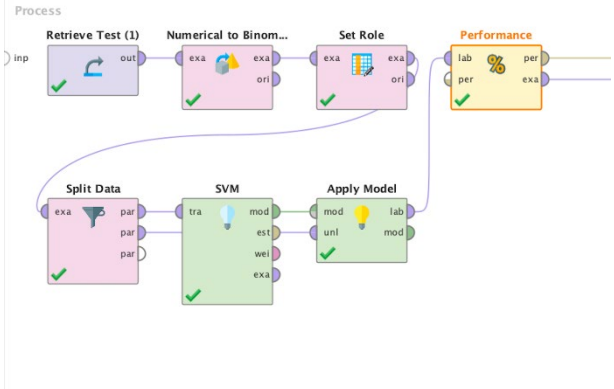

Figure 8: SVM in RapidMiner

4.3. Prediction by using SVM

accuracy: 70.19%

	true false	true true	class precision
pred. false	0	0	0.00%
pred. true	79	186	70.19%
class recall	0.00%	100.00%	

Figure 9: Using rain to predict AQI

accuracy: 70.19%

	true false	true true	class precision
pred. false	0	0	0.00%
pred. true	79	186	70.19%
class recall	0.00%	100.00%	

Figure 10: Using wind speed to predict AQI

accuracy: 70.19%

	true false	true true	class precision
pred. false	0	0	0.00%
pred. true	79	186	70.19%
class recall	0.00%	100.00%	

Figure 11: Using wind direction to predict AQI

accuracy: 70.19%

	true false	true true	class precision
pred. false	0	0	0.00%
pred. true	79	186	70.19%
class recall	0.00%	100.00%	

Figure 12: Using UV ray to predict AQI

accuracy: 70.19%

	true false	true true	class precision
pred. false	0	0	0.00%
pred. true	79	186	70.19%
class recall	0.00%	100.00%	

Figure 13: Using temperature to predict AQI

accuracy: 70.19%

	true false	true true	class precision
pred. false	0	0	0.00%
pred. true	79	186	70.19%
class recall	0.00%	100.00%	

Figure 14: Using humidity to predict AQI

accuracy: 80.75%

	true false	true true	class precision
pred. false	38	10	79.17%
pred. true	41	176	81.11%
class recall	48.10%	94.62%	

Figure 15: Using AQI the day before to predict AQI

5. CONCLUSION AND SUGGESTIONS

5.1. Conclusion

By researching the issue, we first gained a deeper understanding of air pollution and learned many machine learning methods and algorithms. We used one of the methods - SVM to use relevant data to predict AQI values. However, from the operation results it can be seen from the picture that when SVM uses a single factor to predict AQI, the accuracy is only about seventy percent. After discussion, we came to a possible reason: SVM cannot distinguish between right and wrong due to the data, so it guesses all right.

5.2 Suggestion

When collecting data, it's common to encounter significant imbalance between the quantities of positive and negative samples, resulting in what is known as an imbalanced dataset. This imbalance can raise concerns

about the accuracy of predictions for both the majority and minority classes. To address this issue, alternative evaluation metrics are needed to better assess and improve the model's performance.

REFERENCE

Air Quality System (AQS) https://www.epa.gov/aqs
About the Office of Air and Radiation (OAR) https://www.epa.gov/aboutepa/about-office-air-and-radiation
Air Quality index https://airtw.moenv.gov.tw/CHT/Information/Standard/AirQualityIndicator.aspx
Central Weather Administration https://www.cwa.gov.tw/V8/C/D/UVIHistory.html
Ministry of Environment https://data.moenv.gov.tw/dataset/detail/AQX_P_434
Andriy Burkov(2019). The Hundred-Page Machine Learning book. ISBN-13: 978-1999579500
Gaganjot Kaur Kang, Jerry Zeyu Gao, Sen Chiao, Shengqiang Lu, and Gang Xie(2018). Air Quality Prediction: Big Data and Machine Learning Approaches.
Gokulan Ravindirana, Gasim Haydera, Karthick Kanagarathinamd, Avinash Alagumalaie, Christian(2023).Air quality prediction by machine learning models: A predictive study on the indian coastal city of Visakhapatnam .