Air-Pollutant Detection using Support Vector Regression

Aditya Upadhyay **DaxAmin** **Nilay Kapadia**

Priyance Mandlewala **Vaibhava Laxmi** **Vishrut Patel**

ABSTRACT

The increase in air pollution due to fossil fuel consumption as well as its ill effects on the climate has made air pollution forecasting an important research area in today's times. SVR or Support Vector Regression is an effective algorithm, an extension of the highly effective SVM classification algorithm. This algorithm uses the principles of SVM for regression analysis unlike SVM which is used for classification. Moreover, the functional characteristics of the SVM are also investigated in the study. The analysis shown in this report justifies the use case of SVR for air pollutant prediction.

1 INTRODUCTION

Due to the varied topography and extent of industrialization in the US, predicting environment pollutant values help in foreseeing the effect and extent of pollution. The data set provides information on the city, county, NO2, CO, O3, SO2 levels for through 16 years, 2000 – 2016. Relation between the pollutants to their geographical locations translates the problem into a classification issue. Compared to other methods, SVM is particularly useful since the data involves a time series and is non-linearly related. This method can also provide a better generalization error. In order to predict continuous values, however, leads to the use of a variation of SVM - SVR.

2 THEORY

2.1 Support Vector Machines (SVM)

SVMs (Support Vector Machines) are used mainly used for classification problems. The basic idea is to find a hyperplane which separates the d-dimensional data perfectly into two or many classes. However, since the real-world data is often not linearly separable, SVMs introduce the notion of a "kernel induce feature space" which casts the data into a higher dimensional space where the data is separable. Typically, casting into such a space would cause problems computationally, and with over-fitting. The key insight used in SVM's is that the higher-dimensional space doesn't need to be dealt with directly, which eliminates the above concerns. Overall, SMVs are intuitive, theoretically well-founded, and have shown to be practically successful.

We are given *l* training examples $\{x_i,y_i\}$, i=1,...l, where each example has d inputs ($x_i \in \mathbf{R}^d$) are parametrized by a vector (w), and a constant (b) expressed in the equation

$$\boldsymbol{w}.x + b = 0$$

Given such a hyperplane (**w,**b) that seperates the data, this gives the function:

$$f(x) = sign(w.x + b)$$

which correctly classifies the training data. However, a given hyperplane represented by (w,b) is equally expressed by all pairs $\{\lambda w, \lambda b\}$ for $\lambda \in \mathbf{R}^+$. So we define the canonical hyperplane to be that which seperates the data from the hyperplane by a space of atleast 1. That is all the points are considered which satisfy:

$$x_i.w + b \geq \ 1\, when\ y_i = 1$$

$$x_i.w + b \leq \ 1\, when\ y_i = -1$$

2.2 Support Vector Regression

The Support Vector Regression (SVR) uses the same principles as the SVM for classification, with only a few minor differences. First of all, because output is a real number it becomes very difficult to predict the information at hand, which has infinite possibilities. In the case of regression, a margin of tolerance (epsilon) is set in approximation to the SVM which would have already requested from the problem. But besides this fact, there is also a more complicated reason, the algorithm is more complicated therefore to be taken in consideration. However, the main idea is always the same: to minimize error, individualizing the hyperplane which maximizes the margin, keeping in mind that part of the error is tolerated. The model produced by SVR only depends on a subset of the training data, because the cost function for building the model ignores any training data that is close (within a threshold ε) to the model prediction. Suppose we are given training data {(x1, y1),...,(x, y)} ⊂ X × R, where X denotes the space of the input patterns (e.g. X = Rd). In ε-SV regression, our goal is to find a function f (x) that has at most ε deviation from the actually obtained targets yi for all the training data, and at the same time is as flat as possible. In other words, we do not care about errors as long as they are less than ε, but will not accept any deviation larger than this.

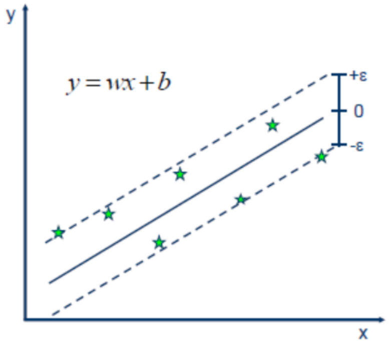

Figure 1: Graph showing Hyperplane for SVR.

Only the points outside the region covered by dotted lines contribute to the cost insofar, as the deviations are penalized in a linear fashion.

3 DATASET AND GRAPHS

3.1 Details about Data

The data used for analysis has been obtained from the US Meteorological department website for the years 2000-2016. It contains statistical data such as mean, max and min for 4 kinds of pollutants namely, Nitrogen Dioxide, Sulphur Dioxide, Ozone and Carbon Monoxide. It also contains temporal data in the form of dates and the data was collected every 6 hours. Moreover, the dataset contains Geospatial data in the form of State code, County Code, Site number, Address of the data collection station and the names of the respective city, county and state.

3.2 Graphs

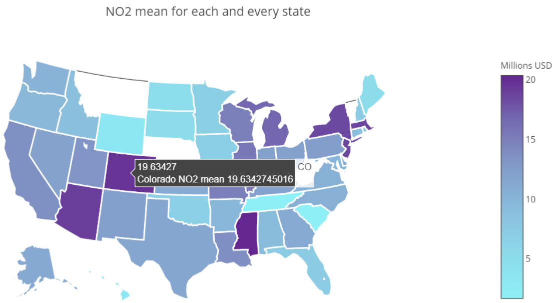
NO2 mean for each and every state
19.63427
Colorado NO2 mean 19.6342745016
CO
Millions USD
20
15
10
5

Figure 2: The distribution of the pollutant NO2 across the United States.

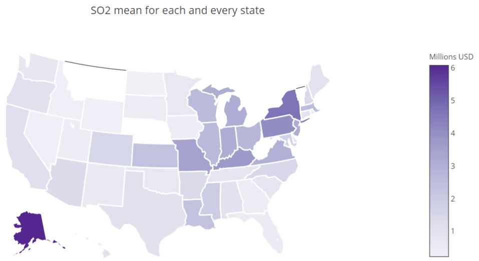

Figure 3: The distribution of the pollutant SO2 across the United States.

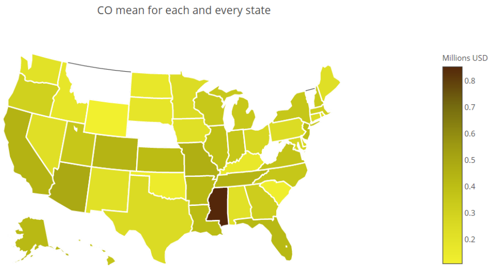

Figure 4: The distribution of the pollutant CO across the United States.

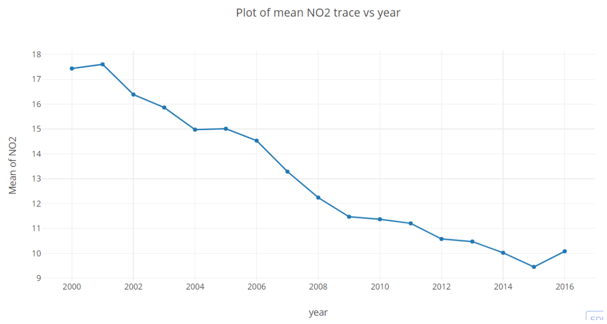

Figure 5: Graph showing the average NO2 concentration over the 16-year period

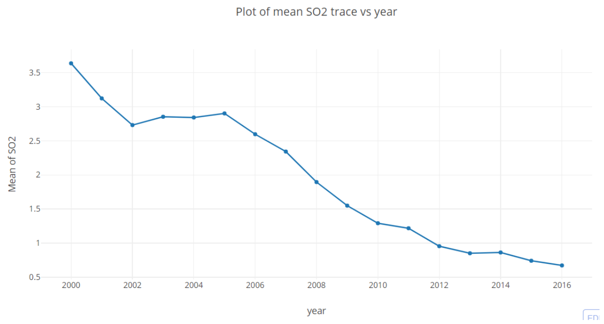

Figure 6: Graph showing the average SO2 concentration over the 16-year period

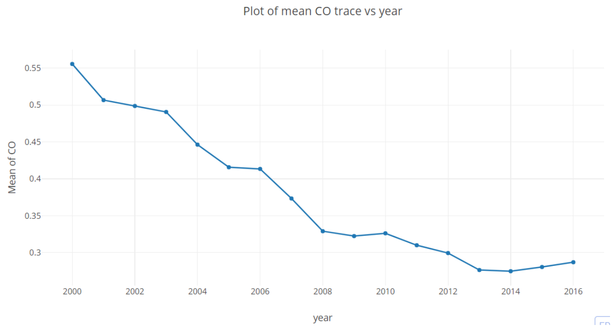

Figure 7: Graph showing the average CO concentration over the 16-year period

4 CONCLUSION

The analysis performed above as well as the statistical output obtained indicates that the use of SVR is a valid approach for the prediction of air pollutant detection. SVR takes into account spatial and temporal data to predict the output as compared to its more primitive version, SVM, which can only be used for classification.

REFERENCES

- Data Source: US Meteorological Department
- SVM: Introduction to Support Vector Machines, Bowell, 2002
- SVR: Support Vector Regression, Basak et.al., 2007