Predicting air pollution level in a specific city

Dan Wei
dan1991@stanford.edu

1. INTRODUCTION

The regulation of air pollutant levels is rapidly becoming one of the most important tasks for the governments of developing countries, especially China. Among the pollutant index, Fine particulate matter (PM2.5) is a significant one because it is a big concern to people's health when its level in the air is relatively high. PM2.5 refers to tiny particles in the air that reduce visibility and cause the air to appear hazy when levels are elevated.

However, the relationships between the concentration of these particles and meteorological and traffic factors are poorly understood. To shed some light on these connections, some of these advanced techniques have been introduced into air quality research. These studies utilized selected techniques, such as Support Vector Machine (SVM) and Neural Network, to predict ambient air pollutant levels based on mostly weather and sometimes traffic variables.

This project attempted to apply some machine learning techniques to predict PM2.5 levels based on a dataset consisting of daily weather and traffic parameters in Beijing, China. Due to the uncertainty of the specific number PM2.5 level, I simplified the problem to be a binary classification one, that is to classify the PM2.5 level into "High" (> 115 ug/m^3) and "low" (<= 115 ug/m^3). The value is chosen based on the Air Quality Level standard in China, which set 115 ug/m^3 to be mild level pollution.

2. DATA OVERVIEW

In order to identify and forecast key parameters affecting air quality and propose appropriate preventive strategies and policies, it is essential to systematically collect data characterizing air quality.

The data includes two parts: training data set and test data set. Training data set has 322 observation points and the test data has 55 points. Each point represents the meteorological and traffic condition of a specific day in Beijing City. The total data set covers 47 days in 2014 and 330 days in 2013.

The data comes from China Meteorological Data Sharing Service System, Beijing Transportation Research Center and US Embassy in Beijing.

As mentioned before, the output data was labeled as one or zero. One refers to high pollution level and zero refers to low pollution level. The total number labeled as zero is 103, while the remaining 274 points are labeled as 0.

3. FEATURE SELECTION

A variety of meteorological, traffic and industrial parameters affect the air pollution level. After taking consideration of the data availability and importance, this project used the following five features:

X_1 - Temperature

Temperature affect air quality because of temperate inversion: the warm air above cooler air acts like a lid, suppressing vertical mixing and trapping the cooler air at the surface. As pollutants from vehicles, fireplaces, and industry are emitted into the air, the inversion traps these pollutants near the ground.

X_2 - Wind speed

Wind speed plays a big role in diluting pollutants. Generally, strong winds disperse pollutants, whereas light winds generally result in stagnant conditions allowing pollutants to build up over an area.

X_3 - Relative Humidity

Humidity could affect the diffusion of contaminant.

X_4 - Traffic index

The large number of cars on the road cause high level of air pollution and traffic jam may increase the pollutants concentration from vehicles. The definition of traffic index is a index reflecting the smooth status of traffic. The index range is from 0 to 10. 0 represents smooth and 10 represents sever traffic jam.

X_5 - Air quality of previous day

The air pollution level is influenced by the condition of the previous day to some extent. If the air pollution level of the previous day is high, the pollutants may stay and affect the following day.

4. METHOD

This prediction is a binary classification problem, so the following three supervised learning algorithms were used:

1) Logistic regression: fitglm

The output is a Generalized Linear Model. For this model, the prediction value is range for 0 to 1. In order to get the label, the values were converted to zero (if $0 \leq$ value ≤ 0.5) and one (if value ≥ 0.5).

2) Naive Bayes Classification: fitcnb

The output is a Classification Naive Bayes classifier.

3) Support Vector Machines: fitcsvm

The output is a Classification SVM classifier. For this model, it was proved that linear Kernel Function gave the best prediction results for this problem.

The models are all from Matlab library.

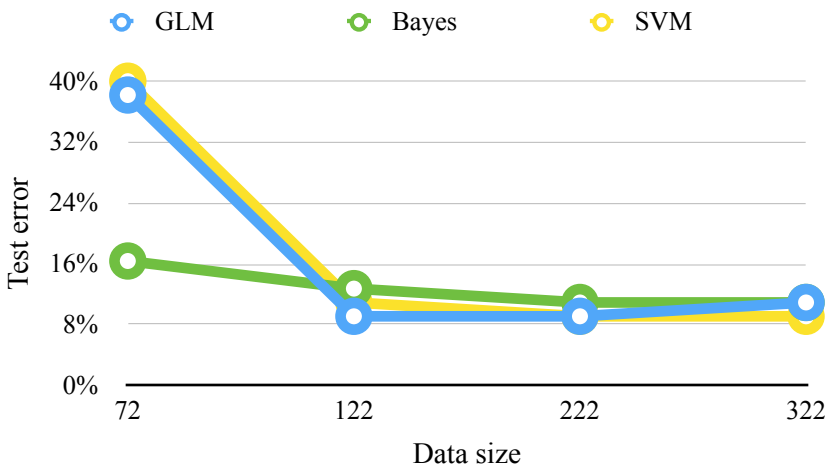
5. RESULT ANALYSIS

1) Error analysis

The total data size is 322. The overall test error for GLM is 10.91%, which is the same as it for Bayes. SVM has the lowest test error, 9.09%.

After changing the data size and repeat training the model, we got the test error curve as shown in figure 1.

Figure 1. Test error curve for three models



The figure 1 shows that in this problem, the test error of Bayes classifier doesn't change much with data size, however GLM and SVM have large test error change with data size. Further more, the test error for SVM has the decline trend if the data size increases further.

2) Prediction performance analysis

Classification-based predictions for test examples can be evaluated using a variety of measures. The most straightforward measure is accuracy, which is the percentage of the examples that are correctly predicted. However, this measure may not be sufficient. This projected chose the measures in Table 1, since they are well understood and have been used extensively in areas such as information retrieval and computational biology, where prediction is a common task.

Table 1. Measures used for evaluating the predictions from the classifiers

Measure	Definition	Notes
Precision (P)	TP/(TP + FP)	For each class, measures how many of the predicted members are actually true members.
Recall (R)	TP/(TP + FN)	For each class, measures how many of the true members are correctly predicted (recovered).
F-Measure	2×P×R/(P+R)	Measures the trade-off between P and R for each class.

TP = no. of true positives, FP = no. of false positives, TN = no. of true negatives, FN = no. of false negatives.

Therefore, the prediction performance for there different models could be evaluated as the summary in Table 2 below.

Table 2. Measures used for evaluating the predictions from the classifiers

Method	Precision (P)	Recall (R)	F-Measure
Logistic regression	0.706	0.923	0.800
Naive Bayes Classification	0.733	0.846	0.785
Support Vector Machines	0.722	1.000	0.839

After training the whole training set, SVM has the highest F-Measure while Naive Bayes has the lowest. This initial result shows that SVM has the overall best performance for predicting the air pollution level in this problem.

6. DISCUSSION

The primary goal of the project was the prediction of air pollution level in Beijing City with the ground data set. The best algorithm (SVM) gave the 0.722 precision, 1.000 recall and 0.839 F-Measure value. It is relatively accurate and is an acceptable result for practical use. However, compared with results from some literatures, the predicting performance (F-Measure value) for this data set is not very good. Also, the advantage of SVM are not shown obviously. It would be better to try other SVM models rather than the one from Matlab.

On the other hand, the data set in this project is not large enough. Air quality is a long-term formed problem and it is better to use a large data data covering a variety of years and locations.

Furthermore, beside the meteorological and traffic factors, industrial parameters such as power plant emissions also play significant roles in air pollution. This project did use these features because they are not public available in China. In order to get better prediction results, the data should include more industrial condition features if possible.

REFERENCES

[1] Pandey, Gaurav, Bin Zhang, and Le Jian. "Predicting submicron air pollution indicators: a machine learning approach." Environmental Science: Processes & Impacts 15.5 (2013): 996-1005.

[2] Athanasiadis, Ioannis N., et al. "Applying machine learning techniques on air quality data for real-time decision support." First international NAISO symposium on information technologies in environmental engineering (ITEE'2003), Gdansk, Poland. 2003.

[3] Ioannis N. Athanasiadis, Kostas D. Karatzas and Pericles A. Mitkas. "Classification techniques for air quality forecasting." Fifth ECAI Workshop on Binding Environmental Sciences and Artificial Intelligence, 17th European Conference on Artificial Intelligence, Riva del Garda, Italy, August 2006.

[4] M. Caselli & L. Trizio & G. de Gennaro & P. Ielpo. "A Simple Feedforward Neural Network for the PM10 Forecasting: Comparison with a Radial Basis Function Network and a Multivariate Linear Regression Model." Water Air Soil Pollut (2009) 201:365–377.

[5] S. Bordignon, C. Gaetan and F. Lisi, "Nonlinear models for ground- level ozone forecasting." Statistical Methods and Applications, 11, 227-246, (2002).