Implementation of Machine Learning Algorithms In Predicting Air Pollution Level Of A City And Their Performance Analysis Thereof

Mohammad Raihanul Bashar, Md. Rakin Sarder
Department of Computer Science & Engineering, Independent University, Bangladesh
1320454@iub.edu.bd
1722341@iub.edu.bd

***Abstract*— The problem of air pollution is a frequently recurring situation and its management has social and economic considerable effects. The goal of this project was to design air quality prediction model using machine learning approach. Dhaka city have been chosen for the project subject city, and air quality data of Dhaka city have been accumulated from air quality monitoring station situated national and international level. Adequate set of per day air quality data of Dhaka was chosen for training, which was then tested with a test dataset for prediction. Three different learning methods have been applied to analyze their performance over the problem scenario, such as: Naïve Bayes, SVM and MLP. The performance analysis of the algorithms in this scenario showed SVM gave the highest F-Measure, including the limitation of the size of the dataset.**

I. INTRODUCTION

The regulation of air pollutant levels is rapidly becoming one of the most important tasks for the governments of developing countries. Air quality forecasting is one of the core elements of contemporary Urban Air Quality Management and Information Systems. Air quality is typically assessed based on either expert meteorologist knowledge or on sophisticated "first principles" mathematical models. Air Quality Operational Centers have been established worldwide in areas with (potential) air pollution problems. These centers monitor critical atmospheric variables and they publish regularly their analysis results [1]. Currently, real-time decisions are made by human experts, whereas mathematical models are used for offline study and understanding of the atmospheric phenomena involved. The goal of this work is real time assessment of air quality. Specific problems in real-time air quality assessment include: sensor malfunction, instrument polarization, noise, etc.

Moreover, rapid environmental changes have rendered previous assessment methods obsolete. At the same time, state regulations worldwide have defined stricter pollution levels. There is a need for new techniques for reliable real-time assessment of air quality based on sampled data. Urban air quality information is created when methods, tools or human judgment is applied over a data set that is usually comprised of time series records resulting from the operation of monitoring stations. Mathematical methods and tools may provide with forecasting capabilities, thus offering decision makers with the opportunity to take preventive measures that would "smooth" or alter the results of a forecasted "episode" or even "crisis". The complexity of air pollution data has been extensively discussed [3], while the usage of various modelling tools is frequently addressed in related literature [4].

In the fields of machine learning and data mining, an extensive arsenal of classification/prediction algorithms has been developed to build models for predicting class labels of examples that are encoded by a set of features (represented by a vector) [5]. Athanasiadis et.al. [6] implemented The σ–FLNMAP classifier which is applicable in a fuzzy lattice data domain including the N-dimensional Euclidean space. The σ–FLNMAP classifier is a synergy of two σ– FLN schemes for clustering. In their work the problem of air quality assessment was addressed in real-time as a classification problem with satisfactory results. Ioannis et. al. [7] evaluated different types of algorithms for classification of air quality. Caselli et.al. [8] used a feed forwarding back-propagation neural network to classify PM 10 data.

II. DATA DESCRIPTION

The dataset contains 576 instances of half hourly response of different air quality parameters for Dhaka, Bangladesh. The data was extracted from aqicn.org and Ministry of Environment and Forests, Govt. of People's Republic of Bangladesh. Aqicn.org is a part of the World Air Quality Index project which is a social enterprise project started in 2007. The project is proving a transparent Air Quality information for more than 70 countries, covering more than 9000 stations in 600 major cities. Their base data source is World Meteorological Organization - surface synoptic observations (WMO-SYNOP), Dhaka Air Quality Monitor - US Consulate, Citizen Weather Observer Program (CWOP/APRS), CPCB - India Central Pollution Control Board and U.S. Embassy and Consulates' Air Quality Monitor in India. Bangladesh Govt. Ministry is running a nationwide project called "CASE", which is involved in air quality monitoring and analysis locally.

Data was extracted from January 1, 2017 to March 5, 2017 (64 days) from aqicn.org and CASE project. The dataset contains half hourly concentrations of PM2.5, PM10, temperature, humidity and wind pressure. This dataset can be used exclusively for research purposes as mentioned specifically by aqicn.org. Commercial purposes are fully excluded.

III. LEARNING METHODS

A. Naïve Bayes

Naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features. Naive Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem. Maximum-likelihood training can be done by evaluating a closed-form expression, which takes linear time, rather than by expensive iterative approximation as used for many other types of classifiers. Given a way to train a naive Bayes classifier from labeled data, it's possible to construct a semi-supervised training algorithm that can learn from a combination of labeled and unlabeled data by running the supervised learning algorithm in a loop:

Given a collection $D = L \cup U$ of labelled samples L and unlabelled samples U, start by training a naive Bayes classifier on L.

Until convergence, do:

Predict class probabilities $P(C \mid x)$ for all examples x in D. Re-train the model based on the probabilities (not the labels) predicted in the previous step.

Convergence is determined based on improvement to the model likelihood $P(D \mid \theta)$,where denotes the parameters of the naive Bayes model. This training algorithm is an instance of the more general expectation–maximization algorithm (EM): the prediction step inside the loop is the E-step of EM, while the re-training of naive Bayes is the M-step.

B. Support Vector Machine

Support vector machines (SVM) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier [9]. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.

Air pollution control is necessary to prevent the situation from worsening in the long run. On the other hand, short-term forecasting of air quality is needed in order to take preventive and evasive action during episodes of airborne pollution. A classical forecasting method is based on multivariate statistical analysis, but now, the artificial neural network (ANN) is becoming an effective and popular means alternatively to conventional methods.

In fact, during the last decade, the increase of computer power has permitted the implementation of many artificial intelligence networks (Hertz et al. 1991; Hecht-Nielsen 1989, 1990; Kohonen 1988; Korn 1991).

The comparison between the computer and the human brain capability provides results dependent on the considered problem. The human brain has some features that would be important to reproduce in the artificial systems.

C. Multi-layer Perceptron

Multi-layer Perceptron (MLP) is a supervised learning algorithm that learns a function $f(\cdot) : R^m \rightarrow R^o$ by training on a dataset, where m is the number of dimensions for input and o is the number of dimensions for output. Given a set of features $X = x_1, x_2, \ldots, x_m$ and a target y, it can learn a non-linear function approximator for either classification or regression. It is different from logistic regression, in that between the input and the output layer, there can be one or more non-linear layers, called hidden layers. Figure 1 shows a one hidden layer MLP with scalar output.

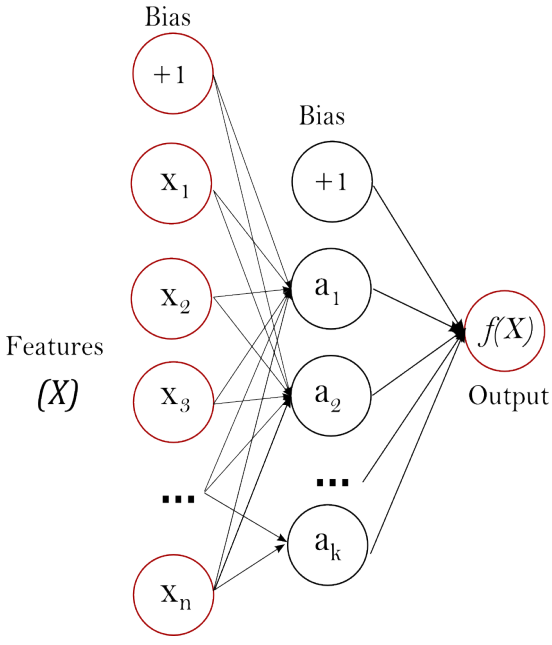

Figure 1: One hidden layer MLP.

The leftmost layer, known as the input layer, consists of a set of neurons $\{x_i | x_1, x_2, ..., x_m\}$ representing the input features. Each neuron in the hidden layer transforms the values from the previous layer with a weighted linear summation $w_1x_1 + w_2x_2 + ... + w_mx_m$, followed by a non-linear activation function $g(\cdot) : R \rightarrow R$ - like the hyperbolic tan function. The output layer receives the values from the last hidden layer and transforms them into output values. MLP trains using Stochastic Gradient Descent, Adam, or L-BFGS. Stochastic Gradient Descent (SGD) updates parameters using the gradient of the loss function with respect to a parameter that needs adaptation, i.e.

$$w \leftarrow w - \eta(\alpha \frac{\partial R(w)}{\partial w} + \frac{\partial Loss}{\partial w})$$

where η is the learning rate which controls the step-size in the parameter space search. $Loss$ is the loss function used for the network.

IV. METHODOLOGY AND ANALYSIS

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

1. Logistic regression: 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: The output is a Classification Naive Bayes classifier.
3. Support Vector Machines: 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 form python library.

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 (fig 2).

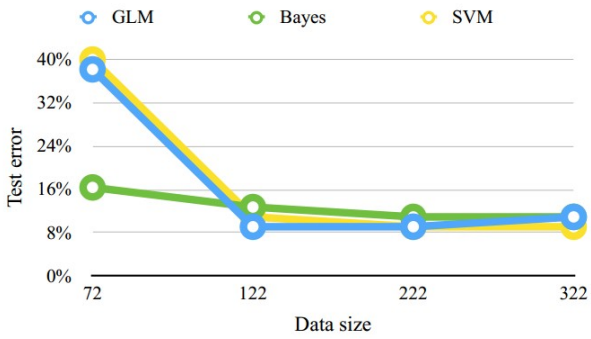

Figure 2: Test error curve of three learning methods for Dhaka Air Quality Data.

The figure 2 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. Furthermore, 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.

Therefore, the prediction performance for three 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
Naïve Bayes Classification	0.733	0.846	0.785
Support Vector Machine	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.

The primary goal of the project was the prediction of air pollution level of a 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. 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 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.

V. CONCLUSION

The primary target was to predict the air pollution level in Dhaka City with the ground data set. The best algorithm (SVM) predict the dataset with 0.722 precision, 1.000 recall and 0.839 F-measure value. It is relatively accurate and is an acceptable output result for practical use. However, comparing with results from other literatures, the predicting performance (F-measure value) for this data set is quite good.

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

Furthermore, beside the traffic and meteorological factors, industrial parameters such as power plant emissions also play significant roles in air pollution. In order to get better prediction and real-time applicable system, the data should include more industrial condition features as possible.

REFERENCES

[1] E. Kalapanides and N. Avouris. Applying Machine Learning Techniques in Air Quality Prediction. In Proc. **ACAI 99**, pp. 58-64, Chania, July 1999

[2] F.M. Morabito and M. Versaci, 'Fuzzy neural identification and forecasting techniques to process experimental urban air pollution data', **Neural Networks**, **16,** 493–506, (2003).

[3] M. Makowski, 'Modeling paradigms applied to the analysis of European air quality, European **Journal of Operational Research**, **122**, 219-241, (2000).

[4] P.-N. Tan, M. Steinbach and V. Kumar, **Introduction to data mining**, Pearson Addison Wesley, Boston, 1st edn, 2006.

[5] 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.

[6] 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.

[7] 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.

[8] Jin, Chi; Wang, Liwei (2012). Dimensionality dependent PAC-Bayes margin bound. Advances in Neural Information Processing Systems.

[9] Hertz, J., Krogh, A., & Palmer, R. G. (1991). Introduction to the theory of neural computation. CA: Addison Wesley.