INTERPOLATION, PREDICTION AND FEATURE ANALYSIS OF FINE-GRAINED AIR QUALITY

Abhijna S Hebbar[1], Aishwarya V[1], Kulsum Begum[1], Namratha A[1], Dr. Rashmi Amardeep[2].

1. *Department of Computer Science Engineering, Sir M Visvesvaraya Institute of Technology, Bangalore, India*
2. *Faculty of Department of Computer Science of Engineering, Sir M Visvesvaraya Institute of Technology, Bangalore, India*

Abstract: The three main topics in the field of metropolitan air quality estimation are interpolation, feature analysis and prediction of the fine-grained air quality. If these topics are resolved, they can provide very useful information to control air pollution and therefore generate great technical and societal impacts. The existing work solves the particular three problems individually by using different models. This project proposes a popular and constructive approach to fix the three issues in a single model. The proposal makes use of the details regarding to the unlabeled, spatiotemporal data to refine the execution of the interpolation, prediction and performs feature selection and associative analysis to disclose the important and suitable features of the quality of the air.

Keywords: Interpolation, Prediction, Feature Selection, Feature Analysis, Spatio-temporal data.

I. INTRODUCTION

Since there are insufficient stations which monitor the air quality, a better interpolation fixes this issue, as the distribution of such stations are also uneven in a city. In order to protect humans from being harmed from air pollution, a precise prediction is necessary to come up with valuable information. The important and suitable factors to the variation of quality of air is disclosed by an appropriate feature analysis. This proposition uses the statistics concerning to the unlabeled, spatio-temporal data in order to upgrade the conduct of interpolation and prediction and selection of features and their analysis. The unlabeled data is utilized here, since it is costlier to acquire the labeled samples due to insufficient stations which monitor the air quality. Also, the labeled samples of data of the stations which monitor the air quality are partial, and there may exist lots of omitted or misplaced labels of the historical data in some stations during some time periods, due to the monitoring devices, as there is a single monitor device per station which has to be sustained at intermissions. Therefore the device does not produce any outputs when it is under maintenance or when it has any problems. Hence, this proposal utilizes the information incorporated in unlabeled and spatio-temporal data and performs the selection of features and their analysis for the metropolitan air related data. Generally, the unlabeled data can assist in delivering suitable information to make the most of the geometric structure of data. The exceedingly valuable information to assist the air pollution control can be extracted by the combined solutions to these topics, therefore provoking prominent technical and societal impressions.

II. RELATED WORK

Lixin Li et al. (2011), have proposed methods of spatiotemporal interpolation for the assessment of air pollution. The accuracies of interpolation results have been compared and the most constructive time scale was chosen to execute the PM2.5 interpolation. They have also assessed the outlook of population to the ambient air pollution of PM2.5 at the county-level.

Yu Zheng et al. (2013), have proposed an approach which provides statistics about the air quality of metropolitan cities. It gives information about two classifiers. One spatial classifier is centered on an artificial neural network (ANN), that takes spatially related statistics to configure the spatial correlation between qualities of air in various locations. Another is a temporal classifier which is centered on a conditional random field (CRF), which takes temporally-related features to configure the temporal dependency of quality of air in a particular location.

Hsieh et al. (2015), have proposed an approach which infers real-time quality of air of any location taking environmental data into account. Also, the method has helped in determining locations to establish new stations to monitor air quality to improve the inference quality. They have proposed a model to minimize entropy which suggests the better locations to establish the new stations to monitor the air quality, based on the metropolitan air data.

Salimol Thomas et al. (2007), have proposed a model to forecast the fine particulate matter(PM2.5) and Carbon Monoxide. They have developed neural network and regression models which helped to forecast the hourly PM2.5 and Carbon Monoxide concentrations. Statistics like the traffic data, time series of particulate matter, were utilized to develop the models. Utilizing many quality indicators these models were then compared. Acceptable accuracy was predicted in the hourly concentration of PM2.5 in both the models.

Ming Li et al. (2015), have proposed an approach to forecast the quality of fine-grained air utilizing the big data. The model forecasts the records of a station that monitors the air quality over the next forty eight hours, by utilizing a data-driven method which takes the present weather forecasts, meteorological data and data of the air quality stations and stations within few kilometers.

Zheng-Jun Zha et al. (2009) have proposed semi-supervised learning methods which are based on graphs, centered on the issue of single label. They have proposed a graph-based learning skeleton of the multi label based semi-supervised learning. The approach is distinguished by parallelly making the most of the inherent correlations between the consistency of labels and multiple labels over the graph.

Avrim Blum et al. (2000), have proposed a model using Co-training to combine the labeled and unlabeled data. When there is an availability of a small set of labeled examples the model uses a large unlabeled sample in order to enhance the performance of an algorithm. They have proposed a Probably Approximately Correct (PAC) framework to find a solution to the problem of using both the unlabeled and labeled data to learn.

Beatriz Maeireizo et al. (2004), have proposed a method to predict the emotions with the data of spoken dialogue using Co-training. The classifiers which predict the emotions in the spoken statements are trained using Co-training. They have applied Naïve Bayes approach for reducing the magnitude of the set of features.

Yingming Liy et al. (2013), have proposed a method which is used to learn using noisy and restricted tagging and have proposed a model to make the most of both the unlabeled and labeled data via a semi-parametric standardization which also makes use of the multi label parameters called SpSVM-MC. This model is centered on the distinct noisy image tagging application using restricted labeled samples on a training dataset.

Richard Socher Jeffrey Pennington et al. (2011), have proposed a model to predict the sentiment distributions using semi-supervised recursive auto encoders. These encoders are constructed using the machine learning frame-work to perform the sentence-level prediction on the collections of the sentiment labels. The multi-word clauses are predicted by using the vector space depictions.

III. METHODOLOGY

A system is developed where the data is loaded, preprocessed and split into train and test data. In addition to this, the system also helps to predict the air quality from the split data sets. The proposed flow diagram is depicted in the figure 1.

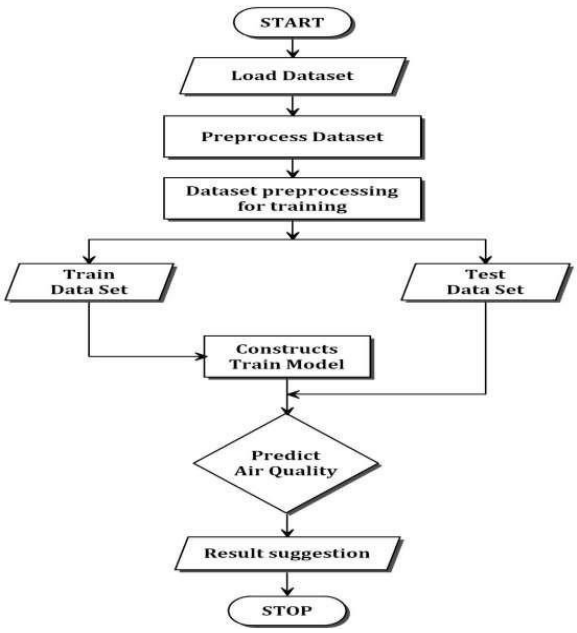

figure 1. flow diagram

The system consists of four stages:

1. Data Acquisition and Preprocessing
2. Feature Selection and Data Preparation
3. Model Construction and Model Training
4. Model Evaluation

I Data Acquisition and Preprocessing:

Data and Models are the two things necessary for the Machine Learning models to work with. The basic data remains in the raw form of digits, qualitative terms and statements. This raw data contains deletions, inaccuracies and inconsistencies. Certain steps are to be taken to scrutinize and process the huge collection of primary data.

The raw data is converted to a clean data set by using the technique called Data Preprocessing, since the real world data consists of:

- Noisy data : The technical issue of a gadget which collects data or a human error while entering the data are some of the reasons for noisy data. This is removed by replacing such values with any suitable values analogous with the other values in data set.
- Inconsistent data : The human data entry, existence of duplication within data, errors in names are some of the reasons for the data to be inconsistent.

- Inaccurate data : The data not collected continuously, biometrics issues, mistakes in data entry are some of the reasons for inaccuracy of data. Such rows are therefore dropped.

II Feature Selection and Data Preparation:

The process where the features that contribute the prediction variable. The technique used here to develop this model is called as Filter-based feature selection technique, which uses mathematical estimates to score the dependence between the variables to be filtered to select the most accurate features.

Classification of data is done to compare the groups of observations, by visualizing the data to find if the training data consists of the target attribute. The dataset is split into two subsets called as training set, used to train the model and testing set, used to test the trained model. Here the ratio of train and test sets is 80:20, done by importing the train_test_split model from the sklearn.model_selection library.

AQI-based data labelling process is used. An air quality index is defined as the overall transformative method in which the weighed parameters (for example, pollutant concentrations) in a single number or collection of numbers are connected to the weighted values of individual air pollution (Ott.1978).The combination function F of individual pollutant sub-indices is prone since most indicators suffer from uncertainty and eclipsing. The maximum operator system, free of both ambiguity and eclipsing as shown below, has been adopted for the proposed AQI: AQI = Max (I1,I2,I3, ... ,In).

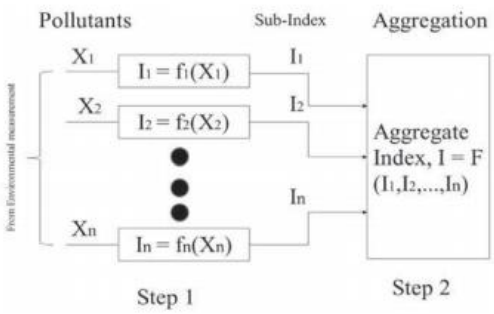

fig 2. formation of an aggregated air quality index

$$AQI = Max(I_1, I_2, I_3, \ldots, I_n)$$

fig 3. AQI maximum function

III MODEL CONSTRUCTION AND TRAINING

The process of coaching an ML model involves providing an ML algorithm (that is, the training algorithm) with training data to find out the results. The term ML model refers to the model artifact that's created by the training process. The training data must contain the right answer, which is understood as a target or target attribute. The learning algorithm finds patterns in the training data that map the input data attributes to the target (the answer that is to be predicted), and it outputs an ML model that captures these patterns.

Our system will be implementing the following algorithms:

- Random Forest
- Naïve Bayes Algorithm
- KNN

I Naive Bayes Classifier

A Naive Bayes Classifier uses the Bayes' Theorem, which assumes that features are statistically independent. The theorem relies on the naive assumption that input variables are independent of each other variable. Regardless of this assumption, it's proven itself to be a classifier with good results.

Naive Bayes Classifiers believe the Bayes' Theorem, which is predicated on contingent probability or in simple terms, the likelihood that an occasion (A) will happen as long as another event (B) has already happened. Essentially, the theory allows a hypothesis to be updated whenever new evidence is introduced.

The equation below expresses Bayes' Theorem :

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

fig 4. naive bayes formula

Where:

"P" is the symbol to denote probability.

P(A | B) = The probability of event A (hypothesis) occurring as long as B (evidence) has occurred.
P(B | A) = The probability of the event B (evidence) occurring as long as A (hypothesis) has occurred.
P(A) = The probability of occurrence of event B. P(B) = The probability of occurrence of event A.

II Random Forest Model

Random forest is a classifier that consists of numerous sub trees. The output of the categories is delineate by individual trees.. This methodology combines with random choice of options to construct a selection trees with controlled variations. Random forest algorithmic rule is a supervised classification algorithmic rule .A random forest is a meta expert that matches the tree classifiers on numerous sub-samples of the dataset and uses averaging to boost the prophetic accuracy and management over-fitting. The sub-sample size is often similar to the original input sample .In the random forest classifier, the higher most tree among the forest provides the high accuracy results.

Random forest algorithm advantages:

• Random forest classifier can handle the missing values.
• Even when we have a lot of trees among the forest, random forest classifier won't over fit the model.
• Will model the random forest classifier for categorical values conjointly.

III K Nearest Neighbors

K nearest neighbors is a easy algorithmic rule that stores all accessible cases and classifies new cases (e.g., distance functions). KNN has been utilized in applied math estimation and pattern recognition from the starting of 1970's as the non-parametric technique.

A case is assessed by a majority vote of its neighbors, with the case being allotted to the class commonest amongst its K nearest neighbors measured by a distance operator. If K = 1, then the case is simply allotted to the class of its nearest neighbor.

KNN is slow supervised learning algorithmic rule, it takes longer to get trained. For classification like alternative algorithmic rule the process is split into 2 step coaching from information and testing it on new instance . The K Nearest Neighbor working principle relies on assignment of weight to the every information that is called as neighbor. In K Nearest Neighbor distance is calculate for coaching information set for each K Nearest data points then the classification is completed on basis of majority of votes.

$$EucledianDistance = Dx, = (xi - yi)\ 2ki = 1$$

K=number of cluster
x, y=co-ordinate sample spaces

$$Manhattandistance = (xi - yi)i{=}1$$

x & y are co-ordinates

$$Min = (-yip)\ 1p$$

Minkowski distances are generally Euclidian distance

Grouping of sample is predicated on super class within the KNN reduction of sample .Selection of k value plays a pivotal role.

The algorithmic rule for KNN is outlined among the steps given below:

1. D represents the samples used within the training and k denotes the quantity of nearest neighbor.
2. Produce super class for every sample category.
3. Calculate Euclidean distance for every coaching sample.
4. Classify the sample, based on majority of classes in neighbor.

IV MODEL VALIDATION

I Analysis and Validation

When it involves examination 2 or additional machine learning algorithmic rule, it's most troublesome as a result of 2 algorithms take issue in some ways. Reason for problem for comparison is , algorithmic rules performance are extremely dependent on the dataset, the only way to decide which algorithm is provides accurate results for the given dataset is by implementing them.

The calculations are made utilizing the information obtained from https://data.gov.in/catalog/historical-daily-ambient-air-quality-data. 80% of the information are used for training and the remaining 20% for testing. Moreover, a number of stages are taken for enhancing the calculations of precision during the process like incorporating cleanup of the dataset and data pre-processing. Calculations are performed for evaluating air quality by applying Naïve Bayes algorithmic rule, KNN algorithmic rule, and Random Forest algorithmic rule. At long last, out of those calculations, best methodology is Random Forest which gives 99% precision.

table 1. compares major machine learning algorithms based on different parameter

Techniques	Outlier	Online Learning	Over fitting and under fitting	Parametric	Accuracy
Naïve Bays	It is less pruned to outline	It can perform on online testing	It does not suffer over fitting and under fitting	It is parametric	High with limited dataset
Random Forest	Outlier does not play critical role in interoperation of dataset by decision tree	It does not supported online learning	It suffers from over fitting and under fitting	Non parametric model	Accuracy depend on the dataset and ensemble technique used.
KNN	It is pruned to outlier	Online learning is supported.	It is more prune to over fitting.	It is parametric	Higher than all other parametric model

table 2. accuracy

	PRECISION	RECALL	FSCORE
NAVIE BAYES	0.8435086786223779	0.8395764727405781	0.8394724720769071
RANDOM FOREST	0.9994709756893895	0.9684184046834979	0.9683599858686341
KNN	0.9683231194928943	0.9684184046834979	0.9683599858686341

V RESULTS

The test is performed for a total of 218638 rows.
Confusion Matrix is obtained after training and testing model to decide if the model is working as expected.
The confusion matrix obtained for Navie Bayes algorithm.

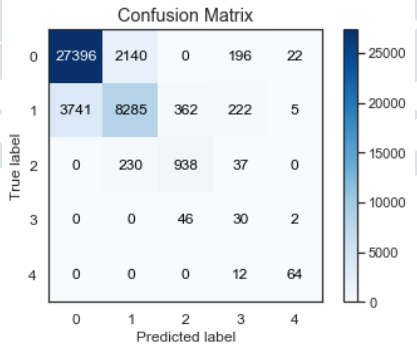

fig 5. confusion matrix-navie bayes

The confusion matrix obtained for Random Forest tree algorithm.

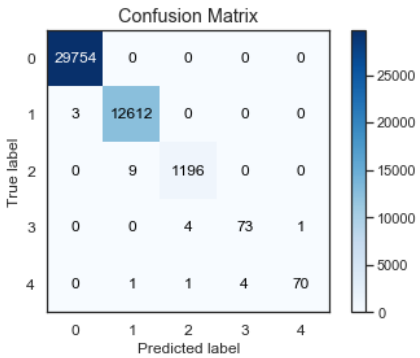

fig 6.confusion matrix-random forest tree

The confusion matrix obtained for KNN algorithm.

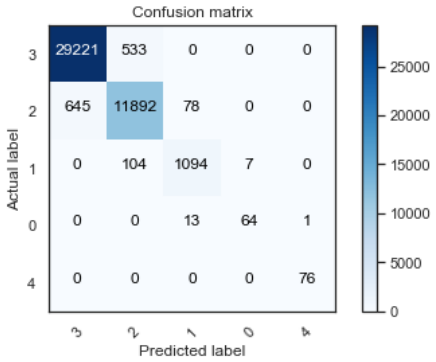

fig 7. confusion matrix- KNN

The analysis of air quality was done better by Random forest tree among the 3 algorithms used which included Navie Bayes, Random Forest tree and K-Nearest neighbors.

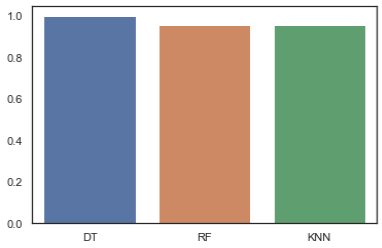

fig 8. comparing results

VI CONCLUSION

Three key topics in the field of urban air computing are: interpolation, prediction and analysis of fine-grained air quality. Such technologies may provide valuable knowledge to promote air pollution management and can have significant social and technological impacts. The various problems are mainly resolved separately by developing different models. We develop in this paper an effective and general approach called DAL to unify the interpolation, prediction, selection of the features and analysis of the quality of fine-grained air into one model. Association analysis combined with feature selection revealed the significance of different inputs to neural network predictions. The proposed selection and analytical function approach helps the deep black box deep network model to expose an internal mechanism. This analysis plays a crucial role in creating a smog warning in the intelligent city and designing environmental policies. With this project based on public data from metropolitan cities, our model predicts the density of smog (air-pollution) gasses in those areas based on combined data from the Historical Daily Ambient Air Quality Data released by the Ministry of the Environment and Forests and Central Pollution Control Board of India, which is largely clean. All this information is accessible on the internet directly or indirectly. In addition, algorithms for other cities can be implemented.

VII ACKNOWLEDGEMENTS

The Author is thankful to the Sir M Visvesvaraya Institute of Technology for providing computer lab facilities.

REFERENCES

[1] “Spatiotemporal Interpolation Methods for Air Pollution Exposure”, Lixin Li, Xingyou Zhang and James B. Holt et al.
[2] “U-Air: When Urban Air Quality Inference Meets Big Data”, Yu Zheng, Furui Liu, HsunPing Hsieh.
[3] “Inferring Air Quality for Station Location Recommendation Based on Urban Big Data”, Hsun-Ping Hsieh, Shou-De Lin, Yu Zheng.
[4] “Model for Forecasting Expressway Fine Particulate Matter and Carbon Monoxide Concentration: Application of Regression And Neural Network Models”, Salimol Thomas & Robert B Jacko.
[5] “Forecasting Fine-Grained Air Quality Based on Big Data”, Yu Zheng, Xiuwen Yi, Ming Li et al.
[6] “Graph-Based Semi-Supervised Learning with Multi-Label”, Zheng-Jun Zha, Tao Mei et al.
[7] “Combining Labelled and Unlabelled Data with Co-Training”, Avrim Blum, Tom Mitchell
[8] “Co-training for Predicting Emotions with Spoken Dialogue Data”, Beatriz Maeireizo, Diane Litman and Rebecca Hwa
[9] “Learning with Limited and Noisy Tagging”, Yingming Liy, Zhongang Qiy, Zhongfei (Mark) Zhang et al.
[10] “Semi-Supervised Recursive Auto encoders for Predicting Sentimen Distributions”, Richard Socher Jeffrey Pennington et al.