ONLINE SCALABLE SVM ENSEMBLE LEARNING METHOD (OSSELM) FOR SPATIO-TEMPORAL AIR POLLUTION ANALYSIS

Shahid Ali and Simon Dacey

Department of Computing, Unitec Institute of Technology, Auckland, New Zealand

ABSTRACT

Environmental air pollution studies fail to consider the fact that air pollution is a spatio-temporal problem. The volume and complexity of the data have created the need to explore various machine learning models, however, those models have advantages and disadvantages when applied to regional air pollution analysis, furthermore, most environmental problems are global distribution problems. This research addressed spatio-temporal problem using decentralized computational technique named Online Scalable SVM Ensemble Learning Method (OSSELM). Evaluation criteria for computational air pollution analysis includes: accuracy, real time & prediction, spatio-temporal and decentralised analysis, we assert that these criteria can be improved using the proposed OSSELM. Special consideration is given to distributed ensemble to resolve spatio-temporal data collection problem (i.e. the data collected from multiple monitoring stations dispersed over a geographical location). Moreover, the experimental results demonstrated that the proposed OSSELM produced impressive results compare to SVM ensemble for air pollution analysis in Auckland region.

KEYWORDS

SVM, Ensemble, Spatio-temporal, Aggregation, Scalable

1. INTRODUCTION

Machine learning algorithms must keep with latest developments in environmental and other applications. Machine learning algorithms performance is dependent on sufficient and reliable data coming from various sources. With internet revolution, nowadays data is easily accessible through electronic sources. However, environmental data is sometimes short and contains missing data because of equipment failure, human error or incorrectly setting up monitoring stations dimensions etc. Therefore, present machine learning models for air pollution prediction analysis do not represent the true picture of air pollution. Further, the existing online models are confronted with the problem of accommodating huge amount of spatiot-emporal data. In this regard scalability of machine learning model and aggregating various model decision plays an important role in online spatio-temporal air pollution prediction. This chapter investigates Online Scalable SVM Ensemble Learning method (OSSELM) towards online spatio-temporal Nitrogen dioxide (NO_2), carbon dioxide (CO) and ozone (O_3) prediction for Auckland over its monitoring stations Auckland wide.

Our daily activities result in release of chemicals and particles in air that we breath. These air pollutants in air can cause hazy days, unpleasant smells and have adverse effects on our health. Air quality is dependent on the amount of pollution released in air by human and natural activities, the degree of diffusion because of wind and weather effects and chemical reaction among the pollutants [1]. The key pollutants in Auckland region that impact the quality of air includes, NO_2, CO and O_3.

In Auckland, New Zealand these pollutants are measured by Auckland Regional Council (ARC) because they are known to endanger human well-being and health. The pollutants concentrations in New Zealand are measured to national standards or to regional air quality targets. On average,

DOI: 10.5121/ijdkp.2017.7602

every Aucklander breathes 11,000 litres of air every day. However, New Zealand's air quality is comparatively good to other nations, but we have the worst asthma death rate [2]. Asthmatics people are sensitive to poor air quality, which in Auckland primarily caused by motor vehicles and other sources such as domestic fires. Hence, the long-term exposure to air pollution results in increase in cardiovascular and lungs diseases resulting in heavy medical costs to public [8] beside damages to other living organisms such as vegetation.

Air pollution threats to public health lead to development of several machine learning models to predict air quality for future. However, air pollution monitoring is a complex task. Monitoring stations consisted of sensor devices which collect each pollutant concentration every minute across whole year. Knowledge discovery from such a large amount of data is again complex, time consuming and quite expensive task. Air pollution which is a spatio-temporal problem, and spatio-temporal data is always in huge amount. Hence, machine learning practitioners are confronted with the problem of handling of this large amount of data and computational time constraint to solve air pollution problem. Yet, machine learning practitioners have to face missing values into the data problem in time-series problem, with missing values a machine learning model would produce biased results towards specific characteristics of spatio-temporal prediction tasks. Hence, keeping the above problems encountered in machine learning, we proposed a method Online scalable SVM ensemble (OSSELM) for air pollution prediction for future. OSSELM will have the ability to handle large amount of spatio-temporal and will have the ability to conduct environmental prediction on missing data. The proposed method will predict air pollution status for whole region based on CO, NO_2 and ground level O_3 concentrations.

Ensembles of SVM models have been used in a wide range of applications such as in medical, engineering and environmental studies [3]. Ensembles are considered as one of the powerful and successful approaches for prediction tasks. There are several studies carried out (e.g. [4] & [5]) to understand ensembles and to enlighten their effectiveness in various applications. Ensemble SVM works on the principle of divide and conquer strategy and various local model's decisions are aggregated through aggregations method to get into the cause of problem. Through sub division strategy training time is heavily reduced resulting in more models to be trained for prediction task. For example, in training of p classifiers on subsamples size of n/p will result in $\Omega(n^2/p)$ approximate complexity. The decreased in complexity will help in dealing with big data sets and nonlinear kernels.

Air pollution data is often time dependent and spatially distributed. Currently there are multiple monitoring stations collecting air pollution data from locally to global geographical scales. Data quality and quantity collected from these multiple monitoring stations depends on the tools used and design of monitoring network etc. In Auckland region, New Zealand CO, NO_2 and O_3 data is monitored extensively by Auckland Regional Council (ARC) consisting of 20 monitoring stations Auckland wide on hourly basis over whole year. However, some of the pollutants monitoring stations were having missing data which created challenge on CO, (NO_2 and O_3 prediction for future air pollution status in Auckland. The data we use of CO, NO_2 and O_3 for prediction consisted of 20 monitoring stations i.e. fifteen urban stations, four rural stations and one industrial station. From the literature, it was quite evident that various traditional prediction models tend to predict air pollution for a specific location, in this chapter we proposed a new decentralised online scalable SVM ensemble method for air pollution prediction for the future state of whole region considering the geographical characteristics of spatio-temporal CO, (NO_2) and (O_3).

The remainder of this research is organized as follows: Section 2 discusses the previous work on online air pollution prediction studies. Section 3 provides methodology for the research.
The information regarding data set, experimental setup and modelling evaluation criteria is provided in section 4. Section 5 focuses on experimental results and discussion and finally in section 6 conclusion to this research is provided.

2. RELATED WORK

In this section, we consider significant works on machine learning and computational methods for air pollution prediction and analysis. The previous work presents a mixture of methods such as neural networks, support vector machines and other approaches towards air pollution forecasting.

Perhaps the earlier work on computational air pollution prediction was started in earliest 19th century. In 1980s a Group Method of Data Handling (GMDH) algorithm was proposed [6]. This algorithm had the ability to not divide the available data into training and testing data for determining the structure of the partial polynomials or determining the number of intermediate variables. The GMDH algorithm was applied for the short term prediction of air pollution concentrations. For this study SO_2 time series data was deployed. Based on the wind velocity and wind direction in Tokushima in Japan a few hours advance SO_2 was developed. The obtained results were compared with a linear regression model and linear autoregressive model, where GMDH algorithm outperformed the other models.

As most of the country population is resided in the cities and their numerous external activities from one place to another result in air population in urban areas. In this regard, a system for monitoring and forecasting air pollution in urban areas was proposed [7]. The proposed system deployed low-cost-air quality monitoring motes that were easily available through an array of gaseous and meteorological sensors. These motes were then transferred to an intelligent sensing platform which consisted of several modules. These modules were responsible for receiving and storage of data and further converting data into useful information for forecasting the pollutants. Three machine learning algorithms such as support vector machines (SVM), artificial neural network (ANN) and M5P model trees were deployed. The results depicted that multivariate modeling with M5P algorithm provided the best forecasting accuracy for SO_2 compare to other to other algorithms.

Nitrogen oxide emission from vehicles emission results in considerable health issues. In this regard, an accurate online support vector regression model (AOVSR) proposed for the emission prediction of nitrogen oxide (NO_x) [9]. It was quite evident from the results that AOVSR performance on small sample data was quite efficient in comparison to support vector regression model and artificial neural network. The proposed model had the capability to predict NO_x emission accurately under certain conditions when parameters were modified. The overall efficiency and prediction accuracy of proposed model was good as the proposed model had the ability to update the parameters by itself with respect to change in time and with change of other parameters.

Particulate matter (PM) is consisted of solid and liquid particles which remains in air for a while, creating a great threat to human health. This provides an opportunity for the researchers to consider causes of PM emission, depending of its level of concentration in air at a certain time and place. Hence, a method for PM emission prediction based on least square support vector machine (LS-SVM) algorithm was proposed [10]. LS-SVM algorithm was based on the principal of reconstruct phase space which was derived from the Takens embedding theorem. In this method, the data was divided into two parts; training and testing. The learning model was obtained by window moving having width n, along the axis time. The results of LS-SVM demonstrated better predication of PM by numerical experiments.

The fast growing of industrial activities resulting in air pollution problem, which is a major concern for public health. An innovative wireless sensor network for air pollution monitoring system (WAPMS) proposed [11]. The proposed system makes use of air quality index for air pollution monitoring in Mauritius. To improve the efficiency of WAPMS a new algorithm for data aggregation named Recursive Quartiles (RCQ) was implemented. This new algorithm RCQ had the ability to eliminate duplicate and invalid readings, which resulted in reduction of data

transmission to centralised station. To handle with any privacy and management issues WAPMS was equipped with hierarchical routing protocol, which caused the motes to sleep in idle time.
It is very important to know the causes of air pollution to avoid further loss to humans and other living organisms. In this regard to know the sources of air pollution principal components analysis was deployed [12]. Along with PCA tree based ensemble learning were constructed for the prediction of urban air quality. It is identified through that PCA vehicles emission and fuel combustion are the main two sources for air pollution presence. Various tree based ensemble learning i.e., decision tree forest, single decision tree and decision treeboost generalization and predictive performance was evaluated and compared with conventional machine learning approaches such as SVM. The miss-classification rate for single decision tree was 8.32%, 4.12% for decision tree forest, 5.62% for decision treeboost and 6.18% for support vector machines. The classification accuracy of decision tree forest and decision treeboost ensembles was comparatively high to classification accuracy of SVM classification and regression, this was successfully done by deploying bagging and boosting algorithms with these trees based ensemble models.

As mentioned earlier that incomplete or missing data results in biased results because of missing data machine learning algorithms confronts the problem of inaccurate prediction performance. In this regard to handle missing environmental data spatial data aided incremental support vector regression (SalncSVR) model for spatio-temporal $PM_{2.5}$ was proposed [13]. In the proposed method, spatial data was used for the training of temporal prediction model. $PM_{2.5}$ data was obtained through 13 monitoring stations of Auckland, New Zealand. The results of SalncSVR model were compared with temporal lncSVR model, where SalncSVR model resulted in better prediction statistics.

Furthermore, some authors have focused their efforts for forecasting air pollution by machine learning approaches such as neural networks, support vector machines and kernel based algorithms. However, to reduce error rate between the model and raw data a mixed approach of consisting support vector machines and kernel functions was proposed for forecasting urban air quality [14]. The kernel functions that were considered for pollutants concentration forecasting of $PM_{2.5,}$ SO_2 and O_3 were consisted of Gaussian, Polynomial and Spline. The application of SVM along with these kernels resulted in good accuracy modelling for pollutant concentration forecasting.

In the past for the effective air pollution forecasting data mining techniques were deployed as well. Artificial neural network model consisting of data mining techniques based on Feed Forward Neural Networks (FFNN) and Multilayer Perceptron (MLP)neural network models were applied for urban and industrial air pollution impacts area [15]. The obtained air pollution patterns shown greater accuracy and lower error rate with MLP neural network model.

Traffic and environmental data is also presented as a time series. Due to dynamic nature of real time data predicting and improving performance of such a task is a great challenge. With such aim a new type of ensembles based on bagging algorithm was proposed to improve the predictive performance of a real time data [16]. Diversity is very important in ensemble creation. In this regard diversity was created through bagged regression tress. However, this study focused on the diversity creation through bagging but fail to highlight the aggregation strategy of decision making of various ensemble models.

Lastly in terms of air pollution daily prediction a method using support vector machines and wavelet decomposition was proposed [17]. The measured time series data was decomposed into wavelet representation and from there wavelet coefficients were predicted. From these wavelet coefficient values the final daily air pollution forecast was prepared. The forecast approach was proposed by applying neural network of SVM type applying a regression mode. However, the study of this work was limited to Gaussian kernel.

From the literature review and to our knowledge we argue the above researches were only attentive on prediction of air pollutant concentrations and to study the impacts of a single pollutant at certain time for research purposes. Spatio-temporal air pollution data is continuously transmitted, streaming asynchronously from multiple locations and hence imposes certain challenges: (i) air pollution problem is spatio-temporal; (ii) dynamic flow of big data streams; (iii) data streams are physically distributed at different places; and (iv) size of the problem could be different, depending on the size of region and number of monitoring stations in that region. Further, it was quite evident from the literature that centralised computing was applied for air pollution analysis. Air pollution data is physically distributed and decentralized monitored through various monitoring stations. Therefore, we strongly argue that centralised computing has certain downsides: (i) centralized data analysis leads to resource challenges, (ii) Online decision making over huge amount of data is difficult, and (iii) system was not scalable, while the problem is scalable.

Hence, to address the above confronts we proposed an Online Scalable SVM Ensemble Learning Method (OSSELM) for air pollution prediction on pollutant concentrations of CO, NO_2 and O_3 for whole region. The main novelties of the proposed OSSELM are: (i) capable of conducting spatio-temporal analysis, (ii) combining distributed monitoring stations knowledge (fusion) through ensemble learning, (iii) online learning enables to accommodate huge amount of data, and (iv) scalability of system enables regional data analysis and knowledge discovery.

3. PROPOSED ONLINE SCALABLE SVM ENSEMBLE LEARNING METHOD (OSSELM)

The common problems associated to solve air pollution problem with centralised computing are: centralised data analysis leads to resource challenges, online decision making to over huge amount of data is difficult and one of the biggest problem of centralised computing is its failure to accommodate large amount of data [22]. All the above challenges lead us for the development of distributed SVM ensemble technique. One of the key contribution of distributed SVM ensemble technique is its cost effective computation and high processing speed for obtaining results. The computation task in this technique is achieved by divide and conquer mechanism, hence reduces the communication load and achieving high computing capacity.

The proposed Online Scalable SVM Ensemble Learning Method (OSSELM) provides solution to spatio-temporal air pollution analysis. OSSELM can handle large amount of regions air pollution data and has the scalability ability to merge knowledge of multiple monitoring stations. Further the proposed method can handle online data streams and system can detect air pollution problem of the whole region and triggering an alarm for potential incident at a certain location. Figure 1 shows the illustration of proposed decnetralised solution for spatio-temporal air pollution analysis.

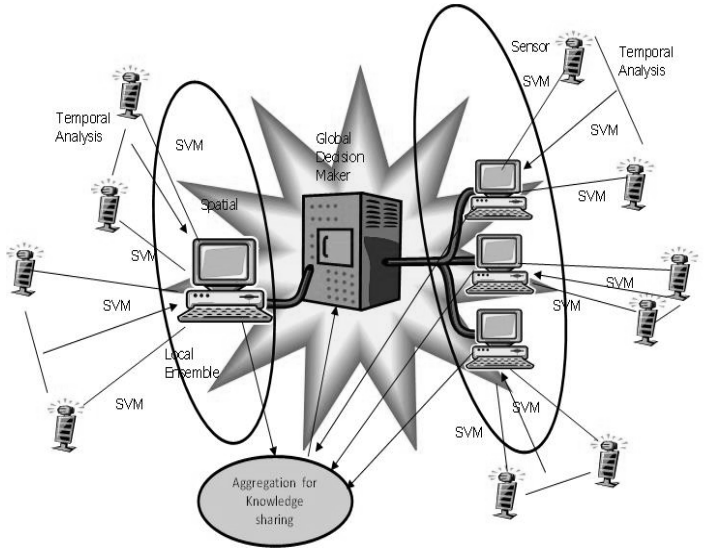

Figure 1. Illustration of Proposed Decentralised Solution for Spatio-temporal Air Pollution Analysis

Figure 1 shows that air pollution analysis is consisted of knowledge from multiple monitoring stations. Each region is composed of various monitoring stations according to the size of region. Based on multiple monitoring stations data SVM ensemble is constructed, which is a collection of numerous SVM ensembles; finally, their knowledge is transferred to the area center for knowledge aggregation and decision support.

3.1 SYSTEM DESIGN

Spatio-temporal air pollution data is obtained from multiple monitoring stations situated in various regions of an area. Local SVMs are modeled on spatial and temporal dimensions. For future decision making regarding air pollution problem knowledge from aspects of various SVMs are integrated. Figure 2 shows the design of OSSELM for spatio-temporal air pollution analysis consisted of four steps.

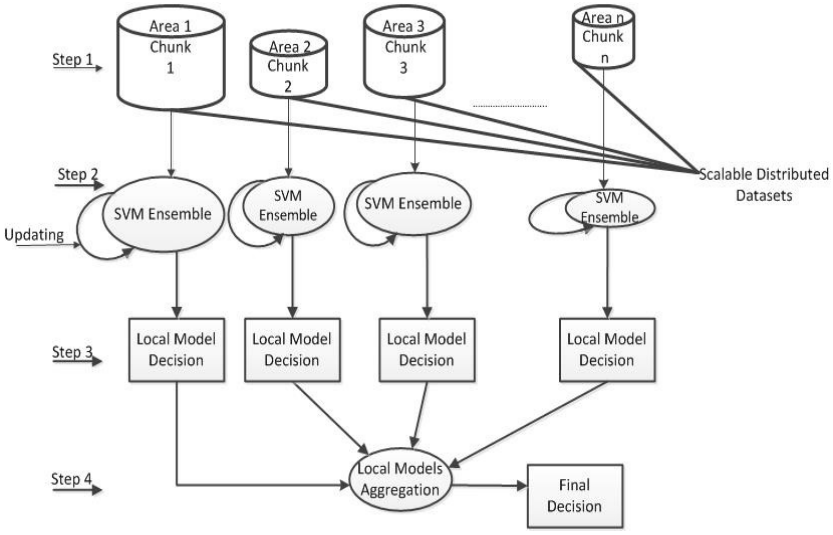

Figure 2. Design of OSSELM for Spatio-temporal Air Pollution Analysis

Step 1: To subsample the region data when it is huge Bag of Little Bootstraps (BLB) is applied.

Step 2: For training of SVM ensembles Boosting method is applied.

Step 3: Multiple individually trained SVM ensembles are combined to build local decision model.

Step 4: Multiple local decision models are aggregated to get the status of air pollution problem for the whole region.

This research follows earlier research and brief detail of above steps can be found in that research having reference [22].

4. DATA SET AND EXPERIMENTAL SETUP

In this section, the proposed OSSELM for air pollution prediction in Auckland region will be experimented on daily hourly CO, NO_2 and O_3 pollutant concentrations collected from 20 monitoring stations in Auckland. Explicitly, we conduct the following cases for: (i) to conduct spatio-temporal prediction for 20 monitoring stations in Auckland; (ii) use ensemble learning to combine multiple stations distributed knowledge; (iii) to accommodate huge amount of data; and (iv) to perform air pollution prediction in real time.

4.1 DATA SET

The data set was obtained from 20 sites comprising of urban, rural and industrial air quality monitoring network of the Auckland Regional Council (ARC), New Zealand. These stations are constantly monitoring air quality pollutants concentrations such as CO, NO_2 and O_3 each hour every day whole year. The CO, NO_2 and O_3 were recorded on an hourly basis 24 hour a day from 1st January 2010 to 31st December 2010. According to the data the recoding of CO, NO_2 and O_3 concentrations, each station has different number of samples depending on the meteorological parameters wind speed, air temperature humidity and evaporation rate.

Table 1 shows air monitoring sites in Auckland region as of April 2010. Table 1 shows that in Auckland region key air pollutants are monitored at various sites to establish the levels of air quality. The current ARC air quality monitoring network comprises of 12 permanent, 1 industrial and 7 mobile sites for pollutant monitoring. All the sites are within the Auckland Urban Airshed. The network extends from Pukekohe in the South to Takapuna in the North, and from Henderson in the West to Botany Downs in the East as shown in Table 1. All sites record CO, NO_2, and O_3. The air quality sites based on 2010 data include: Peak Traffic and CBD sites: Khyber Pass Road (Newmarket) and Queen Street (Auckland CBD) Residential: Dominion Road, Takapuan, Manurewa, Hobson Street, Otara, Henderson, Pakuranga, Dominion road, Newton, Botany Downs, Glen Eden, Saint Marys Bay, Pukekohe Industrial: Penrose (Gavin Street).

Table 1. Air Monitoring Sites in Auckland Region (as of April 2010)

Location	District Council Area	CO	NO_2	O_3
Queen Street III	Auckland	Y	Y	Y
Dominion Road	Auckland	Y	Y	Y
Takapuan	North Shore	Y	Y	Y
Manurewa	Manukau	Y	Y	Y
Hobson Street	Auckland	Y	Y	Y
Khyber Pass Road (A)	Auckland	Y	Y	Y
Henderson	Waitakere	Y	Y	Y
Pakuranga	Manukau	Y	Y	Y
Queen Street II	Auckland	Y	Y	Y
Dominion Road II	Auckland	Y	Y	Y
Newton	Auckland	Y	Y	Y
Botany Downs	Manukau	Y	Y	Y
Penrose IV (A - B)	Auckland	Y	Y	Y
Musick Point	Auckland	Y	Y	Y
Penrose IV (C)	Auckland	Y	Y	Y
Penrose IV (D)	Auckland	Y	Y	Y
Glen Eden	Waitakere	Y	Y	Y
Saint Mary's Bay	Auckland	Y	Y	Y
Pukekohe	Franklin	Y	Y	Y
Number of Sites monitoring Parameters		20	20	20

4.2 EXPERIMENTAL SETUP

The EPI (Environmental Performance Indicators) is a form of Air Quality index (AQI), which are used widely around the world. Currently, EPIs are not widely used and councils now rarely use them. However, back in 2010 Auckland council was using EPIs for monitoring of air pollutant concentrations. The EPI values are different for each air pollutant concentration i.e. CO, NO_2 and O_3, resulting in different classes. The data of this study includes five classes for each pollutant CO, NO_2 and O_3 total of 15 classes for air pollution monitoring in Auckland region. For classification purposes, each pollutant class concentration was represented as with their name along with their level of intensity as shown in Table 2, Table 3 and Table 4 respectively.

Table 2. CO Concentrations and Representations

EPI Class	**Concentration**	**Representation**
Excellent	0-1	CO1
Good	1-3.3	CO2
Acceptable	3.3-6.6	CO3
Alert	6.66-10	CO4
Action	>10	CO5

Table 2 shows the CO concentrations and its representations according to EPI classes. CO concentration between 0 to 1 represented as CO1 and classed as excellent and so on. Whereas,

CO concentration greater than 10 is represented as CO5 and classed as action, meaning immediate action is required to stop further CO pollution spreading.

Table 3. NO_2 Concentrations and Representations

EPI Class	Concentration	Representation
Excellent	0-10	NO1
Good	10-33	NO2
Acceptable	33-66	NO3
Alert	66-100	NO4
Action	>100	NO5

Table 3 shows that NO_2 concentrations and its representations according to EPI classes. NO_2 concentration between 0 to 10 represented as NO1 and classed as excellent and so on. Whereas, NO_2 concentration greater than 100 is represented as N05 and classed as action, meaning immediate action is required to stop further NO pollution spreading.

Table 4. O_3 Concentrations and Representations

EPI Class	Concentration	Representation
Excellent	0-20	OZ1
Good	20-50	OZ2
Acceptable	50-70	0Z3
Alert	70-100	OZ4
Action	>100	OZ5

Table 4 shows that O_3 concentrations and its representations according to EPI classes. O_3 concentration between 0 to 20 represented as OZ1 and classed as excellent and so on. Whereas, O_3 concentration greater than 100 is represented as OZ5 and classed as action, meaning immediate action is required to stop further O_3 pollution spreading.

The data of CO monitoring in Auckland region for this study can be represented as according to EPIs classes: (Class ”CO1”: excellent (meaning, air quality is considered fantastic and no risk at all to people), Class ” CO2”: good (meaning, air quality is considered satisfactory and there is little to people health), Class ” CO3”: acceptable (meaning, air quality is acceptable, however, there is risk to people health), Class ” CO4”: alert (meaning, air quality is not acceptable and there is serious risk to people health), Class ” CO5”: action (meaning, air quality is deteriorating and a quick response is required). NO_2 five classes can be represented as: (Class ”NO1”: excellent (meaning, air quality is considered fantastic and no risk at all to people), Class ” NO2”: good (meaning, air quality is considered satisfactory and there is little to people health), Class ” NO3”: acceptable (meaning, air quality is acceptable, however, there is risk to people health), Class ” NO4”: alert (meaning, air quality is not acceptable and there is serious risk to people health), Class ” NO5”: action (meaning, air quality is deteriorating and a quick response is required). Lastly, O_3 classes can be represented as: (Class ”OZ1”: excellent (meaning, air quality is considered fantastic and no risk at all to people), Class ” OZ2”: good (meaning, air quality is considered satisfactory and there is little to people health), Class ” OZ3”: acceptable (meaning, air quality is acceptable, however, there is risk to people health), Class ” OZ4”: alert (meaning, air quality is not acceptable and there is serious risk to people health), Class ” OZ5”: action (meaning, air quality is deteriorating and a quick response is required).

According to 2010 data set, CO pollution concentrations was recorded from six stations (Takapuna, Khyber Pass Road, Henderson, Pakuranga, Queen Street II and Glen Eden) and consisted of 52560 one hourly observations. NO_2 was recorded from 10 monitoring stations (Khyber Pass Road, Musick Point, Penrose II (B), Takapuna, Henderson, Queen Street II, Glen

Eden, Patumahoe, Waiheke and Hellensville) and consisted of 87600 one hourly observations. O_3 one hourly observations was recorded from four monitoring stations (Patumahoe, Whangaparaoa, Musick Point, Waiheke) and consisted of one hourly 35040 observations.

OSSELM experiments for air pollution prediction on CO, NO_2 and O_3 hourly observations are implemented in Weka 3.6.5 software. We run Weka on Windows 7 Enterprise with system configuration Intel Core i5 processor (3.2GHz) with 4GB 1067 MHz DDR3 of RAM.

4.3 Modelling Performance Evaluation Criteria

We have studied and compare the OSSELM experiments on the parameters like Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Correctly Classified Instances (CCI), Incorrectly Classified Instances (ICI), Percentage Classification Accuracy (PCA) and Kappa Statistics (KS). Mean Absolute Error is used to find out how close the predictions results were to subsequent results. It is calculated based on the average of absolute errors in the predictions. To measure the differences between the values predicted by a model or an estimator to the values observed by changing the values in modelled or estimator are measured through Root Mean Square Error. To measure the inter-agreement of the categorical items Kappa Statistics is used i.e. it is an index which compares the agreement which could happen by chance. Kappa Statistics consists of values ranges from plus one (+1) representing perfectly agreement via zero (0) representing no agreement above that expected by chance, and negative one (-1) representing complete disagreement. This is important to mention here that the prediction performance of OSSELM in our experiments is based on whole data sets of CO, NO_2 and O_3 collected from multiple monitoring stations distributed across Auckland region.

5. Results And Discussions

In this section, we will record proposed OSSELM air pollution prediction performance against SVM ensemble on missing, training, testing and then on unseen data. The OSSELM performance is discussed briefly below.

5.1 OSSELM Performance On Missing And Imputed Missing Data

Air pollution data is mostly confronted with the problem of missing data. As air pollution data is 24/7 collected from multiple monitoring stations at various locations the major reasons for missing values could be equipment failure, power outages, pollutant concentration level is below than detection limits, direction of monitoring station or because of some other unknown reasons [16]. Computational air pollution studies on missing data will not give us the comprehensive knowledge of underlying problem. We argue that any computational method designed for air pollution analysis and prediction should perform similarly on missing and complete data set. In this regard, we test the performance of OSSELM method on missing data of CO, NO_2 and O_3 and its results were compared with SVM ensemble. The comparative statistics of SVM ensemble and OSSELM performances on missing data are shown in Table 5 and Table 6 respectively.

Table 5. SVM Ensemble Performance on Missing Data

Bagging						AdaboostM1					
MAE	RMSE	KS	ICI	CCI	PCA	MAE	RMSE	KS	ICI	CCI	PCA
0.092	0.215	0.252	102138	73061	41.701	0.029	0.169	0.252	3644	138785	79.215

Table 5 shows the SVM ensemble performance on missing data with bagging algorithm resulted in percentage of classification accuracy (PCA) of 41.0701%. Mean absolute error recorded as 0.092 and root mean square error was 0.215. The kappa statistics value was 0.252. SVM ensemble performance on missing data with adboostM1 algorithm resulted in percentage of

classification accuracy (PCA) of 79.215%. Mean absolute error recorded as 0.029 and root mean square error was 0.215. The kappa statistics value was 0.252. Table.1.6 shows the OSSELM performance on missing data with bagging algorithm resulted in percentage of classification accuracy (PCA) of 79.215%. Mean absolute error recorded as 0.029 and root mean square error was 0.169. The kappa statistics value was 0.739. OSSELM performance on missing data with adboostM1 algorithm resulted in percentage of classification accuracy (PCA) of 78.88%. Mean absolute error recorded as 0.093 and root mean square error was 0.190. The kappa statistics value was 0.734.

To solve air pollution prediction problem multiple learners are trained in ensemble learning (EL). Ensemble contains various base learners which results in much stronger generalization ability. Bagging and boosting algorithms are implemented in SVM ensemble for spatio-temporal air pollution with base classifiers such as Single Decision Tree (SDT) and Decision Stump (DS) with their parameters. SDT along with bagging is used for classification of air pollutants and has many features such as to exclude insignificant features, ability to deal with collinear and unbalanced data [17]. DS along with boosting is used for classification of air pollutants and has a number features such producing a decision tree with one single split, further the resulting tree can be deployed to classify unseen data [10]. Various SVM ensemble model decisions on various data chunk are aggregated by averaging method for final decision.

Table 6. OSSELM Performance on Missing Data

Bagging						AdaboostM1					
MAE	RMSE	KS	ICI	CCI	PCA	MAE	RMSE	KS	ICI	CCI	PCA
0.029	0.169	0.739	36414	138785	79.215	0.093	0.190	0.734	36999	138200	78.88

In Table 6 SVM ensemble model with bagging algorithm and SDT as a base classifier provides the air pollution prediction model on missing data resulting in 41.701 percentage of accuracy. SVM ensemble provides the high number of ICI even more than the CCI. This shows that SVM ensemble is not suitable to conduct prediction on missing data or due to missing data SVM ensemble performance is reduced. Furthermore, our previous experiments on SVM ensemble for Spatio-temporal air pollution prediction in chapter 5 resulted that SVM ensemble with bagging algorithm suitable when the number of observations is not too large, for this experiment the number of observations is 175199. The Kappa Statistic value for SVM ensemble on bagging algorithm is far from 1 (i.e. 0.2521), which indicates that SVM ensemble with bagging algorithm along with SDT does not provide the perfect agreement for classification on air pollution pollutants concentrations. SVM ensemble based on adaBoostM1 algorithm provides better results on missing air pollution data compare to bagging having high percentage model accuracy of correctly classified instances of 79.215 percentage.

OSSELM consists of bagging and adaboostM1 algorithms for spatio-temporal air pollution prediction with base classifiers such as REP Tree (RT), Random Tree (RandT), NB Tree (NT), J48, Decision Stump (DS) along with their parameters. In OSSELM base classifiers such as Random Tree (RandT), REP Tree (RT) and NB Tree (NT) are used with bagging algorithm for air pollution prediction and OSSELM performance based on bagging algorithm is recorded. RANDT is extensively used for classification, it takes the input feature vector and classifies it with the tree in forest and outputs the class label that has the majority of vote [18]. RT is an ensemble learning classifier and creates various individual learners consisted of multiple trees and selects the best tree from all the trees generated. It uses bagging concept and produces random set of data to construct decision tree [19]. NB is considered as best classification method in prediction tasks [20].

Further, in OSSELM base classifiers such as J48, Decision Stump and NB Tree are used with adBoostM1 algorithm for air pollution prediction (do you want to mention here that best results

were obtained these classifiers) and OSSELM performance based on adaBoostM1 algorithm is recorded. J48 is a predictive base classifier, it targets the value of a new sample based on the discrete attribute values of the given sample data. One of the key benefit of using J48 classifier is that it cut down the search time for the sorted elements [21]. Various SVM ensemble model decisions in OSSELM on various data chunk are aggregated by majority of voting method for final decision.

From Table 6 comparing the performance of OSSELM with SVM ensemble we can say the OSSELM prediction based on bagging algorithm performed well having accuracy of 79.215 percentage compare to SVM ensemble which is 41.701 percentage. However, OSSELM prediction based on bagging and adaBoostM1 algorithms is considerably same on missing air pollution data i.e.79.215 percentage and 78.88 respectively. It was noticed that SVM ensemble based on adaBoostM1 algorithms resulted in better model accuracy of 79.215 percentage, almost same accuracy of 79.215 percentage and 78.88 respectively with OSSELM with bagging and adBoostM1 algorithms. However, it is noticeable that the Kappa Statistics value for SVM ensemble based on bagging and adaBoostM1 algorithms is 0.2521 compare to 0.7349 of OSSELM with bagging and adaBoostM1 algorithms, it indicates that the value of OSSELM is much closer to 1 (0.7349), which further indicates that perfect agreement for classification of data items is only possible with OSSELM. The above descriptive evaluation statistics Figure 3 shows that the performance of OSSELM on missing data is noticeable enhanced the accuracy of air pollution prediction on various pollutants classes and performed relatively better than SVM ensemble.

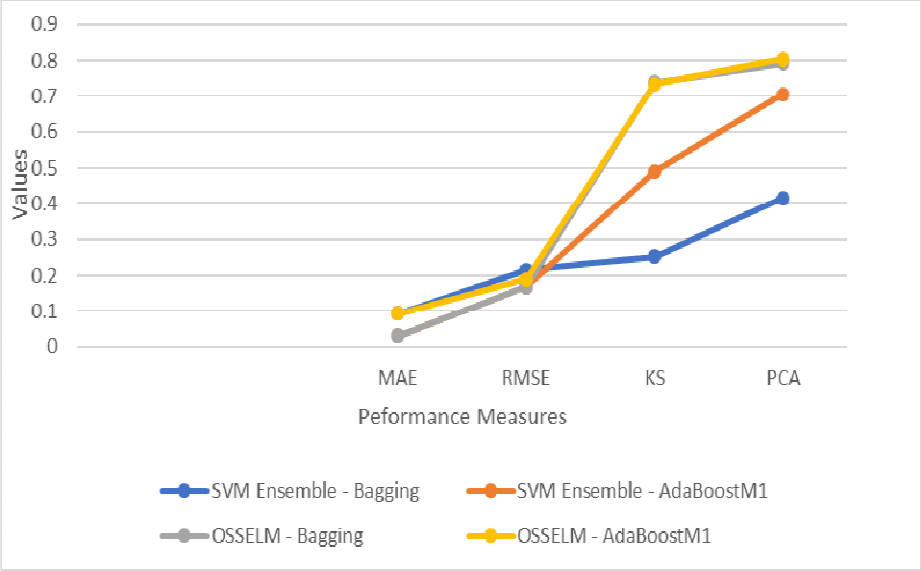

Figure 3. OSSELM vs SVM Ensemble Performance on Missing data

To further validate OSSELM prediction accuracy and other evaluation measures, we used Series Mean Method (SMM) for imputation of missing data values for air pollution concentrations of CO, NO_2 and O_3 as shown in Table 7.

Table 7. Imputation of Missing data with Series Mean Method (SMM)

Pollutants	No of Obs	Missing obs	Missing values (%)	ObsMean	ObsMed	Obs St.d	Imputed Missing obs	Imputed data (%)	Imputed Mean	Imputed. Med	Imputed St.d
NO_2	87600	3845	4.4	16.642	8.627	19.007	3845	4.4	16.64	9.5	18.585
CO	52560	1467	2.8	0.492	0.281	0.599	1471	2.8	0.492	0.3	0.591
O3	35040	726	2.1	43.529	44.77	15.556	726	2.1	43.529	44.3	15.394

Table 7 is about the imputation of missing data with series mean method (SMM). Table.1.7 shows the missing values of air pollutant concentrations which are imputed with SMM. The percentage of missing values for CO, NO_2 and O_3 recorded as 2.8 percent, 4.4 percent and 2.1 percent respectively. The observed mean (ObsMean), observed median (ObsMed) and observed standard deviation are 16.642, 8.627 and 19.007 respectively. The imputed values for CO, NO_2 and O_3 recorded slightly changes in the mean, median and standard deviation indicating the imputed pollutant concentrations values recorded well by the SMM. In this section, we highlight the performance evaluation of SVM ensemble and OSSELM on missing data, in the next section we will assess OSSELM and SVM ensemble model prediction performances on imputed missing data of CO, NO_2 and O_3.

The available data of CO, NO_2 and O_3 consisted of 175200 observations. The observations are separated into training and testing data by using filters in Weka software. The size of the training data is set to 70 percent i.e. 122640 observations of the available data, whilst the test set contained 30 percent i.e. 52560 of the observations. The proposed OSSELM and SVM ensemble are trained and tested using the same exact data to allow for paired comparisons.

5.2 OSSELM Performance On Training Data

The comparative statistics of SVM ensemble and OSSELM on training data are shown in Table 8 and Table 9 respectively.

Table 8. SVM Ensemble Performance on Training Data

Bagging						AdaboostM1					
MAE	RMSE	KS	ICI	CCI	PCA	MAE	RMSE	KS	ICI	CCI	PCA
0.275	0.368	0.495	35560	87080	71.004	0.348	0.389	0.49	35961	86678	70.677

Table 8 shows the SVM ensemble performance on training data with bagging algorithm resulted in percentage of classification accuracy (PCA) of 71.004%. Mean absolute error recorded as 0.275 and root mean square error was 0.368. The kappa statistics value was 0.495. SVM ensemble performance on training data with adboostM1 algorithm resulted in percentage of classification accuracy (PCA) of 70.677%. Mean absolute error recorded as 0.348 and root mean square error was 0.389. The kappa statistics value was 0.49.

Table 9. OSSELM Performance on Training Data

Bagging						AdaboostM1					
MAE	RMSE	KS	ICI	CCI	PCA	MAE	RMSE	KS	ICI	CCI	PCA
0.129	0.348	0.688	23807	98832	80.587	0.194	0.298	0.687	23845	98794	80.556

Table 9 shows the OSSELM performance on training data with bagging algorithm resulted in percentage of classification accuracy (PCA) of 80.587%. Mean absolute error recorded as 0.129

and root mean square error was 0.348. The kappa statistics value was 0.688. OSSELM performance on training data with adboostM1 algorithm resulted in percentage of classification accuracy (PCA) of 80.556%. Mean absolute error recorded as 0.194 and root mean square error was 0.298. The kappa statistics value was 0.687.

SVM ensemble with bagging algorithm provided the highest percentage of classification accuracy i.e. 71.004 compare to SVM ensemble with adaBoostM1 algorithm of 70.6774 percent. Experiment results show that OSSELM provided the highest percentage of classification accuracy both with bagging and adaBoostM1 algorithms i.e. 80.587 and 80.556 respectively.

The number of incorrectly identified instances is high in SVM ensemble both with bagging and adaBoostM1 algorithms resulting in lower value of Kappa Statistic value (i.e. 0.495 and 0.490 respectively). However, OSSELM comparatively perform well to SVM ensemble resulting in lower number of incorrectly identified instances. The Kappa Statistic value for OSSELM for bagging and adaBoostM1 is closer to 1 (i.e. 0.688 and 0.687 respectively) which indicates OSSELM resulted in perfect agreement for classification of various classes of CO, NO_2 and O_3 concentrations.

Mean absolute error for SVM ensemble with bagging is lower compare with SVM ensemble with adaBoostM1 algorithm i.e. 0.275 and 0.348 respectively. Comparing mean absolute values of OSSELM with bagging and adaBoostM1 algorithms resulted in lowest i.e. 0.129 and 0.194 respectively. This indicates that OSSELM with bagging and adaBoostM1 algorithms provided closer predictions for air pollution observations and resulted in lower root mean square error to SVM ensemble.

The above performance evaluation results clearly indicate as shown in Figure 4 that OSSELM outperform SVM ensemble on statistical ground resulting in better prediction results for our analysis. Now we will conduct experiments on the unseen testing data and will record statistics for both OSSELM and SVM ensemble.

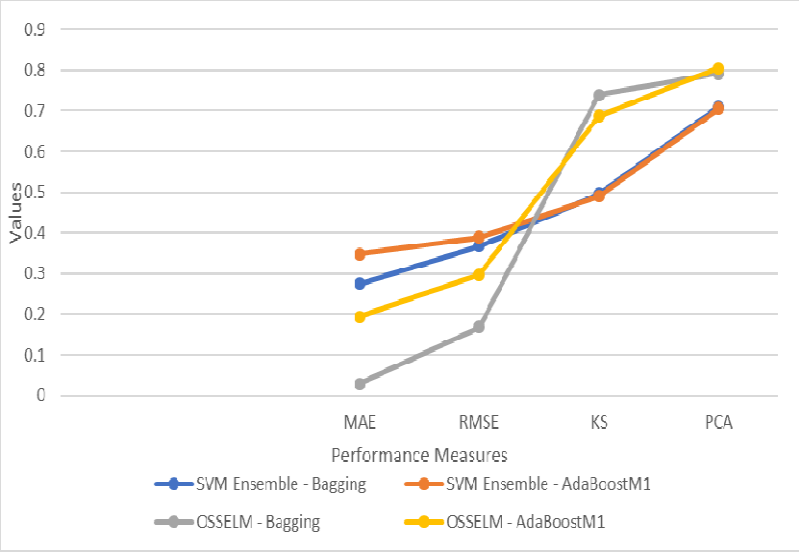

Figure 4. OSSELM vs SVM Ensemble Performance on Training data

5.3 OSSELM Performance On Testing Data

The testing data of CO, NO_2 and O_3 consisted of 52560 observations i.e. 30 percent of 175200 observations. The prediction performance of OSSELM and SVM ensemble is tested on these

unseen observations. The comparative statistics of OSSELM and SVM ensemble on testing data are shown in Table 10 and Table 11.

Table 10. SVM Ensemble Performance on Testing Data

Bagging						AdaboostM1					
MAE	RMSE	KS	ICI	CCI	PCA	MAE	RMSE	KS	ICI	CCI	PCA
0.270	0.366	0.490	15052	37508	71.362	0.270	0.366	0.499	15052	37509	71.362

Table 10 shows the SVM ensemble performance on testing data with bagging algorithm resulted in percentage of classification accuracy (PCA) of 71.362%. Mean absolute error recorded as 0.270 and root mean square error was 0.366. The kappa statistics value was 0.490. SVM ensemble performance on testing data with adboostM1 algorithm resulted in percentage of classification accuracy (PCA) of 71.362%. Mean absolute error recorded as 0.270 and root mean square error was 0.366. The kappa statistics value was 0.499.

Table 11. OSSELM Performance on Testing Data

Bagging						AdaboostM1					
MAE	RMSE	KS	ICI	CCI	PCA	MAE	RMSE	KS	ICI	CCI	PCA
0.131	0.337	0.687	10158	42403	80.673	0.193	0.298	0.683	10274	42287	80.453

Table 11 shows the OSSELM performance on testing data with bagging algorithm resulted in percentage of classification accuracy (PCA) of 80.673%. Mean absolute error recorded as 0.131 and root mean square error was 0.337. The kappa statistics value was 0.490. OSSELM performance on training data with adboostM1 algorithm resulted in percentage of classification accuracy (PCA) of 80.453%. Mean absolute error recorded as 0.193 and root mean square error was 0.298. The kappa statistics value was 0.683.

Table 10 shows that SVM ensemble percentage of accuracy on testing data with bagging and adaBoostM1 algorithms is same i.e.71.362 respectively. How- ever, OSSELM percentage of accuracy is recorded consistent as 80.673 and 80.453 with bagging and adaBoostM1 algorithms as shown in Table 11, which is better than SVM ensemble performance. Kappa statistic value for SVM ensemble with bagging and adaBoostM1 algorithms for both recorded as 0.499, resulting in far from 1. Whereas, OSSELM value for Kappa Statistic with bagging and adaBoostM1 algorithm is recorded as 0.687 and 0.683 respectively, which is closer to 1 resulting in perfect agreement for classification of various classes of CO, NO_2 and O_3 concentrations. OSSELM values for mean absolute error recorded lowest i.e. 0.131 and 0.193 respectively, for bagging and adaBoostM1 algorithms in contrast to SVM ensemble values. Hence, OSSELM with bagging and adaBoostM1 algorithms provided closer predictions for air pollution observations. Similarly, OSSELM values for root mean square error recorded lowest to SVM ensemble.

Based on the above statistics results described Figure 5 clearly indicates that OSSELM outperform SVM ensemble in classification accuracy and other statistical performance measures. The proposed OSSELM successfully predicted the various classes of CO, NO_2 and O_3.

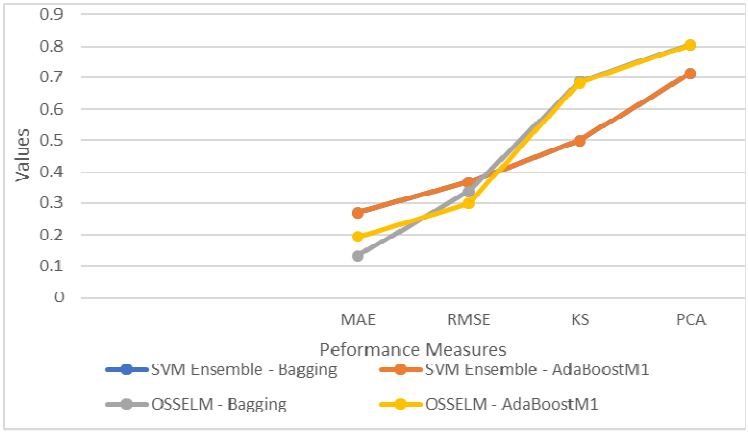

Figure 5. OSSELM vs SVM Ensemble Performance

6. CONCLUSION

Air pollution data is often collected from multiple monitoring stations located at various regions. The size of data of each monitoring stations varies. Air pollution prediction for a data which is distributed across various regions is a challenging task. The proposed OSSELM has the potential of merging knowledge of multiple air pollution monitoring stations distributed across various regions. Local SVMs are modelled on spatial and temporal dimensions, respectively. Further, for future air pollution prediction, the knowledge from various aspects of SVMs is aggregated. The problem of spatio-temporal air pollution prediction based on CO, NO_2 and O_3 is deployed with OSSELM and SVM ensemble method and their performances are recorded. The overall results of OSSELM are shown in Table 12 and the overall SVM ensemble results are shown in Table 13.

Table 12. Overall OSSELM Performance

	Bagging			AdaBoostM1		
	Missing	Training	Testing	Missing	Training	Testing
MAE	0.029	0.129	0.131	0.093	0.194	0.193
RMSE	0.169	0.348	0.337	0.19	0.298	0.298
KS	0.739	0.688	0.687	0.734	0.687	0.683
PCA	79.215	80.587	80.673	78.881	80.556	80.453

Table 13. Overall SVM Ensemble Performance

	Bagging			AdaBoostM1		
	Missing	Training	Testing	Missing	Training	Testing
MAE	0.092	0.275	0.27	0.029	0.348	0.270
RMSE	0.215	0.368	0.366	0.169	0.389	0.366
KS	0.252	0.495	0.499	0.2521	0.49	0.499
PCA	41.701	71.004	71.362	79.215	70.677	71.362

By comparing the overall results of OSSELM in Table 12 with overall results of SVM ensemble in Table 13 it is quite evident that the proposed OSSELM has strong capability of classifying spatio-temporal air pollution problem with missing, training and testing data of various stations with bagging and adboostM1 algorithms.

This eliminates the limitation of some of the proposed models in spatio-temporal domains where missing data creates a question mark on the validity of the performance of model. The proposed OSSELM performance recorded same with missing and imputed missing data. The main objectives of the proposed model achieved by; (i) successfully constructing ensemble learning based classification of various classes of CO, NO_2 and O_3; (ii) scalable SVM models based on ensemble learning are developed and their performances are evaluated statistically and compared with SVM ensemble approach as a benchmark, (iii) the proposed OSSELM resulted in compressive air pollution prediction for whole region, which in our case is Auckland, moreover, most of the environmental authorities i.e. Auckland Regional Council (ARC) and The National Institute of Water and Atmospheric Research (NIWA), are interested in knowing the air quality status in whole region not to a specific location. With this we can conclude, that the proposed OSSELM performed well in different conditions i.e. with missing data and without missing data, and can be used by environmental authorities as a tool in air quality prediction and further analysis.

ACKNOWLEDGEMENTS

The authors would like to thanks Sreenivas Sremath Tirumala for his technical support.

REFERENCES

[1] Xue, Y., Tian, H., Yan, J., Zhou, Z., Wang, J., Nie, L., . . . others (2016). Temporal trends and spatial variation characteristics of primary air pollutants emissions from coal-fired industrial boilers in beijing, china. Environmental Pollution, 213, 717-726.

[2] Soeren Mattke, P. S. J. H. M. L. G. L., Edward Kelley. (2006). Health care quality indicators project: Initial indicators report, oecd health working papers no. 22. Organisation for Economic Co-operation and Development (OECD), 1-152.

[3] Wang, Q., & Zhang, L. (2010). Ensemble learning based on multi-task class labels. In Advances in knowledge discovery and data mining (Vol. 6119, p. 464-475).

[4] Dietterich, T. (2000). Ensemble methods in machine learning. In Multiple classifier systems (Vol. 1857, p. 1-15). Springer Berlin/Heidelberg.

[5] Brown, G. (2009). Ensemble learning.

[4] Tamura, H., & Kondo, T. (1980). Heuristics free group method of data handling algorithm of generating optimal partial polynomials with application to air pollution prediction. International Journal of Systems Science, 11(9), 1095-1111.

[5] Shaban, K. B., Kadri, A., & Rezk, E. (2016, April). Urban air pollution monitoring system with forecasting models. IEEE Sensors Journal, 16(8), 2598-2606.

[6] Perera, F. P. (2017). Multiple threats to child health from fossil fuel combustion: Impacts of air pollution and climate change. Environmental Health Perspectives, 125(2), 141.

[7] Zhou, J., Ji, Y., Qiao, Z., Si, F., & Xu, Z. (2013, July). Nitrogen oxide emission modeling for boiler combustion using accurate online support vector regression. In 2013 10th international conference on fuzzy systems and knowledge discovery (fskd) (p. 989-993).

[8] H. Li, Z., & Yang, J. (2010, July). Pm-25 forecasting use reconstruct phase space ls-svm. In 2010 the 2nd conference on environmental science and information application technology (Vol. 1, p. 143-146).

[9] Khedo, K. K., Perseedoss, R., Mungur, A., et al. (2010). A wireless sensor network air pollution monitoring system. arXiv preprint arXiv:1005.1737.

[10] Singh, K. P., Gupta, S., & Rai, P. (2013). Identifying pollution sources and predicting urban air quality using ensemble learning methods. Atmospheric Environment, 80, 426–437.

[11] Song, L., Pang, S., Longley, I., Olivares, G., & Sarrafzadeh, A. (2014, July). Spatiotemporal pm2.5 prediction by spatial data aided incremental support vector regression. In 2014 international joint conference on neural networks (ijcnn) (p. 623630).

[12] Sotomayor-Olmedo, A., Aceves-Fernandez,´ M. A., Gorrostieta-Hurtado, E., Pedraza-Ortega, C., Ramos-Arregu´ın, J. M., & Vargas-Soto, J. E. (2013). Forecast urban air pollution in mexico city by using support vector machines: a kernel performance approach.

[13] Christy, S., & Khanaa, V. (2016, February). Impacts of ambient air quality data analysis in urban and industrial area helps in policy making. International Journal on Recent Innovation Trends in Computing and Communication, 153-157.

[14] Oliveira, M., & Torgo, L. (2014). Ensembles for time series forecasting. In Acml.

[15] Osowski, S., & Garanty, K. (2007). Forecasting of the daily meteorological pollution using wavelets and support vector machine. Engineering Applications of Artificial Intelligence, 20(6), 745-755.

[16] Carslaw, D. C., & Ropkins, K. (2012). Openair an r package for air quality data analysis. Environmental Modelling & Software, 27, 52-61.

[17] Coops, N. C., Waring, R. H., Beier, C., Roy-Jauvin, R., & Wang, T. (2011). Modeling the occurrence of 15 coniferous tree species throughout the pacific northwest of north america using a hybrid approach of a generic process-based growth model and decision tree analysis. Applied Vegetation Science, 14(3), 402-414.

[18] Sewaiwar, P., & Verma, K. K. (2015). Comparative study of various decision tree classification algorithm using weka.

[19] Kalmegh, S. (2015). Analysis of weka data mining algorithm reptree, simple cart and random tree for classification of indian news. International Journal of Innovative Science, Engineering and Technology, 2(2), 438-46.

[20] Vadhanam, B. R. J., Mohan, S., Ramalingam, V., & Sugumaran, V. (2016). Performance comparison of various decision tree algorithms for classification of advertisement and non advertisement videos. Indian Journal of Science and Technology, 9(48).

[21] Luan, C.,& Dong, G. (2017). Experimental identification of hard data sets for classification and feature selection methods with insights on method selection. arXivpreprint arXiv:1703.08283.

[22] Ali, S., Tirumala, S. S., & Sarrafzadeh, A. (2014, Dec). Svm aggregation modelling for spatio-temporal air pollution analysis. In 17th ieee international multi topic conference 2014 (p. 249-254).

AUTHORS

Shahid Ali

Shahid obtained his Master in Computer Sciences from the Auckland University of Technology, Auckland in 2006. He got industry experience as an IT Analyst and Business Analyst. He is a part time lecturer in Computer Sciences Department at Unitec. He is in his final year Doctorate having thesis entitled, “SVM Aggregation modeling for spatio-temporal air pollution analysis”. He also hold some Microsoft and Cisco certifications i.e. MCP, MCSA, MCSE, CCNA, CCNP, CCDP. http://dmli.info/index.php/member.html

Simon Dacey

Dr. Simon Dacey is full time Computer Sciences lecturer at Unitec Institute of Technology and worked in software development for 17 years on applications as diverse as a firing control system for anti-submarine warfare and a database application for the management of wetland resources in Indonesia. He obtained his MSc in Applied Remote Sensing from the University of Cranfield, UK (1995). From 1998 to 2002 he worked as a lecturer at UCOL in Palmerston North, North Zealand. Since January 2002, he has been a full-time lecturer at the Department of Computing, Unitec Institute of Technology. His research interests include Geographical Information Systems, Remote Sensing, Digital Image Processing, Geographical Positioning Systems, and Database Management Systems.
http://www.unitec.ac.nz/about-us/contact-us/staff-directory/simon-dacey