Semi-supervised Hybrid Modeling of Atmospheric Pollution in Urban Centers

Ilias Bougoudis, Konstantinos Demertzis, Lazaros Iliadis,
Vardis-Dimitris Anezakis, and Antonios Papaleonidas(✉)

Democritus University of Thrace,
193 Pandazidou st., 68200 N Orestiada, Greece
ibougoudis@yahoo.gr,
{kdemertz,liliadis,danezaki}@fmenr.duth.gr,
antonis.pap@gmail.com

Abstract. Air pollution is directly linked with the development of technology and science, the progress of which besides significant benefits to mankind it also has adverse effects on the environment and hence on human health. The problem has begun to take worrying proportions especially in large urban centers, where 60,000 deaths are reported each year in Europe's towns and 3,000,000 worldwide, due to long-term air pollution exposure (exposure of the European Agency for the Environment http://www.eea.europa.eu/). In this paper we propose a novel and flexible hybrid machine learning system that combines Semi-Supervised Classification and Semi-Supervised Clustering, in order to realize prediction of air pollutants outliers and to study the conditions that favor their high concentration.

Keywords: Pollution of the atmosphere · Air quality · Semi-supervised learning · Semi-supervised clustering · Semi-supervised classification · Air pollution

1 Introduction

1.1 Contamination of the Atmosphere

Air pollution is the presence of air pollutants in quantity, concentration or duration, which can cause deterioration of the structure, composition and characteristics of the atmospheric air. The main sources of air pollution are associated with human activities and they are mainly located in urban areas. They are associated with the production of energy, transport, industry and the heating of buildings, engineering structures and households. Air pollution can cause serious health, environmental, social and economic problems. It is caused mainly from oxides, such as oxides of nitrogen, sulfur carbon and soot (unburnt carbon in air mixture gases). Nitrogen oxides cause photochemical smog, usually in cities or centers and the surrounding areas. Oxides of sulfur and carbon react with water vapor cloud creating acid rain, which affects forests, while the sulfuric acid (component of acid rain) attack the marble transforming them into plaster. Carbon dioxide and other gases produced by incomplete combustion, such as unburned

C. Jayne and L. Iliadis (Eds.): EANN 2016, CCIS 629, pp. 51–63, 2016.
DOI: 10.1007/978-3-319-44188-7_4

hydrocarbons, contribute to the greenhouse effect. There are many respiratory events and lung cancer cases in cities that are close to power plants that burn fossil fuels such as oil or lignite. The European Union has announced a strategy aimed at improving the legislation on air quality, in establishing maximum risk limits for various pollutants. The final target is the progressive drastic reduction of emissions, in order to achieve lower morbidity and mortality as a result air pollution. There are primary pollutants emitted directly into the air (e.g. CO, NO, NO_2, SO_2) and secondary formed by chemical reactions between primary ones (e.g. O_3). Although the atmosphere has physicochemical mechanisms that can remove air pollutants, pollution incidents are mainly due to “unfavorable” weather conditions that significantly limit this potential of the atmosphere and they act in a catalytic manner.

Sunshine helps catalyze the transformation of primary pollutants in secondary, speed and direction of the wind influence the dispersion and transport of the pollutants, the stability of the atmosphere due to excitation of the pressure gradient and temperature gradient of the atmosphere also affects the transport and dispersion of pollutants. Moisture creates the effect of atmospheric water vapor. Moreover, the combination of temperature and humidity (Discomfort Index) aggravates the consequences in people with respiratory or heart problems. To make a quantitative assessment of the impact, especially in densely populated urban areas, requires a detailed spatiotemporal analysis of the conditions that favor high pollutant concentrations, focusing on flexible and realistic modeling approaches. Passive monitoring is one of the traditional ways of coping with this phenomenon, without serious substantial forecasting or early intervention and prevention policies. Real-time monitoring and forecasting of pollutants’ concentrations, based on advanced machine learning approaches is one of the most important issues of modern environmental science and research. This research proposes an innovative and effective hybrid forecasting system that does not require high computational power. It employs Semi Supervised Clustering and Classification in order to determine the most extreme air pollutants’ values in urban areas.

1.2 Literature Review – Advantages of the Proposed System

In an earlier research of our team [1] we have made an effort to get a clear and comprehensive view of air quality in the center of the city of Athens and also in the wider Attica basin. This study was based on data that were selected from nine air pollution measuring stations during the temporal periods (2000–2004, 2005–2008 and 2009–2012). This method was based on the development of 117 partial ANN whose performance was averaged by using an ensemble learning approach. The system used also fuzzy logic in order to forecast more efficiently the concentration of each feature. The results showed that this approach outperforms the other five ensemble methods. Also, in a previous research effort, Iliadis et al. [2] applied Self Organizing Maps (SOM) in order to cluster air pollution concentrations in groups. The ultimate goal was to find the most isolated cluster where all of the extreme values of pollutants were gathered. This specific cluster would contain vital information about the hazardous pollutants and would also specify the meteorological and temporal conditions under which they occur. Moreover, they tried to evaluate the clustering outcome, using

Pattern Recognition. The inputs were related to 5 temporal parameters, 7 meteorological and 5 pollutants. Bougoudis et al. [3] present the EHF forecasting system which allows the prediction of extreme air pollutant values. EHF was introduced and tested with a vast volume of actual data records. Its main advantage is that though it takes no pollutants as inputs it manages to operate quite efficiently. Moreover, it used a small number of inputs (7), which comprised of 4 temporal features, air temperature, a station identification code (which was determined automatically by geolocation based services) and a cluster identification code. Four unsupervised learning algorithms were employed in EHF, namely: SOM, Neural Gas ANN, Fuzzy C-Means and a fully unsupervised SOM algorithm. Every algorithm, aimed in detecting the most extreme clusters, which contained the most hazardous pollutants' values. Thereafter, they gathered all the records from the extreme clusters, in order to create four datasets, one for each algorithm. These four datasets were used as inputs to the EHF model, which has given promising results in forecasting pollutants' concentrations.

There are other similar studies in the literature that are trying to forecast the air pollution values. However, they have certain limitations that do not guarantee their generalization ability. More specifically they train ANN models with data related to a narrow area (e.g. city center) and they consider this data sample as representative of a wider area that covers locations varying from a topographic, micro climate or population density point of view. However, such research efforts [4–7] are quite interesting and they offer motivation to scientists from diverse fields to employ artificial intelligence in air pollution modeling. Also there are important seasonal studies in the literature [8–13] that do not offer more generalized annual models. Finally, a very interesting approach with objective criteria has been proposed for the specific problem in China [14]. Also Vong et al. [15] have built a forecasting system based on Support vector machines (SVMs), Xiao et al. [16] proposes a novel hybrid model combining air mass trajectory analysis and wavelet transformation to improve the artificial neural network (ANN) forecast accuracy of daily average concentrations of $PM_{2.5}$ and Zabkar and Cemas [17] have applied methods of machine learning to the problem of ground level ozone forecasting. This requires the use of actual raw data and data calculated by the numerical weather prediction model or stations. On the other hand, Lopez-Rubio et al. [18] introduced Bregman divergences in self-organizing models, which are based on stochastic approximation principles, so that more general distortion measures can be employed. A procedure is derived to compare the performance of networks using different divergences. Moreover, a probabilistic interpretation of the model is provided, which enables its use as a Bayesian classifier. Experimental results show the advantages of these divergences with respect to the classical Euclidean distance. Also Menéndez et al. [19] proposed a new algorithm, named genetic graph-based clustering (GGC), which takes an evolutionary approach introducing a genetic algorithm (GA) to cluster the similarity graph. The experimental validation shows that GGC increases robustness of spectral clustering and has competitive performance in comparison with classical clustering methods. Donos et al. [20] have presented a study to provide a seizure detection algorithm that is relatively simple to implement on a microcontroller, so it can be used for an implantable closed loop stimulation device. The classification of the features is performed using a random forest classifier. Finally, Quirós et al. [21] have extended the traditional definitions of k-anonymity, l-diversity and t-closeness of

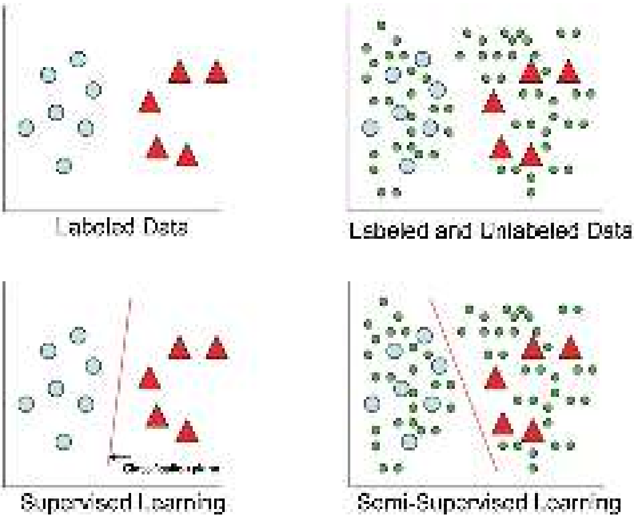

Fig. 1. Semi-supervised versus supervised learning

fuzzy sets as a way to improve the protection of privacy in microdata. The performance of these new approaches is checked in terms of the risk index. The methodology described herein is an extension of a previous research effort of our team [4]. More specifically, [4] describes a system that performs fast and reliable air pollution modeling in mobile devices with limited computational resources. The Semi-Supervised system described in this paper, manages to perform effective air pollution modeling (available to the public) with the minimum amount of data. Unlike paper [4] that uses data input from all the measuring stations in Attica, the proposed system herein uses only data from Athens city center "Athena station". The resulting model has shown high generalization ability for the whole city. Moreover, the system proposed in this research uses fewer features than the initial one presented in [4]. The main advantage of the approach discussed here is its portability. Due to its low requirements and to its generalization potential it can be used in many other cities with the same problem.

The main drawback of the classic classification methods with full supervision (Supervised approach) is that they require a vast number of labeled training examples to construct a predictive model with satisfactory accuracy. The classification of the training set is usually done manually by the instructor, which is a tedious and time consuming process. Instead, the key characteristic of training with partial supervision (Semi-Supervised method) is the production of the final model with the use of pre-classified along with unsorted examples. The Semi-Supervised Clustering approach operates on the condition that the input patterns with and without data tags, belong to the same marginal distribution, or they follow a common cluster structure. Generally, unclassified data provide useful information for the exploration of the overall dataset data structure, while respectively the sorted data are offering in the learning process. Overall, it should be stressed that the success of learning with partial supervision (Semi-Supervised Learning) depends on some basic assumptions imposed by each model or algorithm. This fact makes each case depending on these assumptions which are related to the logic of machine learning methods. Thus, even the most serious real world problems can be modeled effectively, based on the essential peculiarities that characterize them (Fig. 1).

This research effort proposes a Semi-Supervised Classification and Semi-Supervised Clustering Hybrid Air Quality system (SSC^2-HAQS). The system is capable of modeling air pollution in urban centers, after considering the actual positive or negative correlations between all of the involved features (meteorological or primary and secondary chemicals).

2 Data

The data used come from the "Athena" station. The station is located in the heart of Athens, so it provides a representative picture of the atmospheric pollution in modern cities. There were hourly data values available for CO, NO, NO_2, O_3 and SO_2, measured in μg/m^3 for the period 2000–2013. The model was built with data for 13 years (2000–2012) whereas the dataset of 2013 was used for testing the forecast framework with first time seen cases and to determine its validity.

Apart from the five pollutants, each record also consists of five calendar items namely: Year, Month, Day, D_ID (1 Monday, 2 for Tuesday and so on), Hour. Moreover, there are seven meteorological factors namely: Air Temperature (Temp), Relative Humidity (RH), Atmospheric Pressure (PR), Solar Radiation (except 2013) (SR), Percentage of Sunshine (till 2010) (SUN), Wind Speed (WS) Wind Direction (WD). The following tables present a typical statistical analysis of the whole available data and for the 2012 dataset which was chosen to be the pilot one for the determination of the classes (Tables 1 and 2).

Table 1. Statistical analysis for the period 2000–2013

2000–2013 (97201)	CO	NO	NO_2	O_3	SO_2
MAX	21.4	908	377	253	259
MIN	0.1	1	1	1	2
MODE	0.8	7	60	3	2
COUNT_MODE	5592	2651	1606	7137	10435
AVERAGE	1.79	57.88	61.86	33.16	9.40
STANDARD_DEV	1.45	88.29	26.98	28.47	9.06

3 Description of the SSC^2-HAQS Algorithm

We considered three situations for each entry: Tag (1) for the cases with extreme primary pollutants, (2) for records with extreme ozone values (secondary pollutant) and (0) for normal pollutant values. This was assumption was done in order to classify our data in three basic risk categories. Then from the set of available data for 2012, we have

Table 2. Statistical analysis for the year 2012

2012 (8644)	CO	NO	NO_2	O_3	SO_2
MAX	9.3	600	142	186	47
MIN	0.2	1	5	1	2
MODE	0.7	8	53	2	4
COUNT_MODE	672	299	206	436	1372
AVERAGE	1.29	42.36	51.11	38.29	6.88
STANDARD_DEV	0.91	59.67	19.05	29.33	3.28

chosen a small sample of approximately 10 %, which had records that could be clearly labeled as members of one of the three classes. This small sample was used as a pilot in order to classify the rest of the data by employing the Naïve Bayesian algorithm described below.

```
Algorithm 1. The semi supervised Naive Bayesian clustering
Inputs: Input data, clusters of the input data and testing data to which a label should
be assigned
    Step 1:
        Identify the discrete number of clusters
        For every cluster, create matrices with the mean and standard deviation of all
their input data
    Step 2:
        For every cluster, recreate these matrices, based on the testing data
        Calculate a variable, based on the formula below:
 x =(1./(2*pi*ns.^2)).*exp(-((test-nm).^2)./(2.*sn.^2))
        where ns is the new standard deviation matrix, nm is the new mean matrix
and test is the
        testing data
        Sum all these variables for each cluster
    Step 3:
        For every testing data, find the maximum value of the summary calculated
before.
```

Once completing the clustering with the use of the Naive Bayesian algorithm, we have managed to obtain a clear view for the risk level of each record. The corresponding class was added as a new attribute to the final dataset. However, the values assigned to the “0” label were of no interest because the main target was the determination of the extreme cases, regardless the normal ones. The addition of this feature has ensured uniformity as to the classification of the cases and it has solved the following problem:

The concentrations of O_3 in many cases appear to be extremely high, whereas at the same time the relative concentrations of CO and No appear to be extremely low and vice versa. Thus, an overall risk index for both the primary and the secondary pollutants is not possible. The final version of the dataset includes as independent variables the time profile (Year, Month, Day, Day_Id, Hour), meteorological indications (Air-Temp, RH, PR, SR, WS, WD) and the value of the cluster determination to which each record belongs (Cluster). The five pollutants (CO, NO, NO_2, O_3, SO_2) were used as dependent variables.

Then, the Yatsi algorithm was used to classify the unlabeled data, using the classified 10 % as a pilot model. It should be mentioned that the Yatsi algorithm is semi-supervised and it applies the Weighted Nearest Neighbor approach.

Collective classification [22] is a combinatorial optimization problem, in which we are given a set of nodes, $V = \{V1, \ldots, Vn\}$ and a neighborhood function N, where $Ni \subseteq V \setminus \{Vi\}$. Each node in V is a random variable that can take a value from an appropriate domain. V is further divided into two sets of nodes: X, the nodes for which we know the correct values (observed variables) and, Y, the nodes whose values need to be determined. The actual task is to label the nodes $Yi \in Y$ with one of a small number of labels, $L = \{L1, \ldots, Lq\}$; The lower case yi will be used to denote the label of node Yi.

Algorithm 2. High level pseudo code for the two-stage Yatsi algorithm [23]

```
Input: a set of labeled data Dl and a set of unlabeled data Du, an of-the-shelf
classifier C and a nearest neighbor number K; let N = |Dl| and M = |Du|
Step 1:
    Train the classifier C using Dl to produce the model Ml
    Use the model Ml to "pre-label" all the examples from Du
    Assign weights of 1.0 to every example in Dl
    and of $F \times N/M$ to all the examples in Du
    Merge the two sets Dl and Du into D
Step 2:
    For every example that needs a prediction:
    Find the K-nearest neighbors to the example from D to produce set NN
    For each class:
    Sum the weights of the examples from NN that belong to that class
    Predict the class with the largest sum of weights.
```

In this research effort, semi-supervised classification has been applied to isolate the potential extreme records. The reasoning of the method is based on the concept that performing classification for a robust subset of the available data (not less than 10 % of the whole) can provide a prototype for the effective classification throughout the dataset. The following specific steps were applied to achieve this task:

We have initially determined the actual three risk classes, working with the 2012 dataset (pilot data). This was chosen as the actual robust dataset, because it is an extensive one (951 vectors) with the vast majority of the selected values being valid. Also the range of the values for each involved feature was representative of the total potential fluctuation for each pollutant.

Thus it was determined that all 2012 vectors that had CO concentration higher than 3.2 mg/m^3 were labeled as class 1, whereas the ones that had $O_3 > 60\ \mu g/m^3$ were tagged as class 2. All of the rest of the cases were assigned class 0. The above boundary values were selected to represent the extreme cases, based on the results emerging from a previous research effort of our team [3]. Also it is really important that in [2] we had shown that a record can be an outlier, either according to the concentration of primary pollutants (CO, NO, NO_2, SO_2) or based on the secondary O_3 concentrations but not for both types of features at the same time. The parameter CO was selected as representative of the extreme pollutant group 1, because according to [2] it played the most crucial role for its determination, with the extreme values of the rest of primary pollutants to "follow". So we adopted three risk classes for each record: The extreme one for the primary pollutants (1), the extreme in relation to ozone (2) (secondary pollutant) and the class of normal pollutants' values (0). Running the semi-supervised algorithm, we obtained a very effective classification for the whole available data records related to all of the years under study (Table 3).

Correctly Classified Instances 935 (98.3176 %), Incorrectly Classified Instances 16 (1.6824 %), Root Mean Squared Error 0.1024. Figure 2 presents a graph of the proposed method.

After the classification, a dataset with the extreme values of the pollutants was developed. Also the class attribute was added, having the corresponding values 0, 1, 2. This addition ensured uniformity as to the classification of the cases, which appear to have inverse effects over periods of time due to their physico-chemical composition. For example, in cases where there were O_3 outliers, the values of the primary pollutants CO and NO appeared to be extremely low and vice versa.

We have developed feed forward Artificial Neural Networks (ANN) in order to forecast the extreme values of pollutants. Specifically, for each pollutant an ANN has been developed. The input parameters are the following: YEAR, MONTH, DAY, HOUR, AIR_TEMPERATURE and finally the attribute produced by the SSC^2-HAQS, CLUSTER_ID. The network had 10 neurons in the hidden layer, it employed the tansig transfer function, the training function trainlm and the learngdm learning function. The Root Mean Square Error metric (RMSE) was used to evaluate the performance.

4 Results and Comparative Analysis

4.1 Results

Thus, after running the SSC^2-HAQS approach in order to obtain the extreme dataset and after having generated a neural network for each pollutant, the testing process considered the 2013 data vectors, originating from the "Athenas" station. The ANN were not fed with the desired output pollutants' values (targets). Using first time seen inputs, the models predicted some values which were compared to the actual ones (Table 4). The following Table 5 contains the ANN testing results.

4.2 Comparative Analysis

The application of the semi supervised algorithm gives reliable results, especially in the classification process. More specifically, the SSC^2-HAQS model outperforms the approaches that have been proposed by our research team in the literature [3].

Table 3. Confusion matrix for the assignment of the classes

Confusion matrixs 951 instances (0 normal values) (1 extreme primary) (2 extreme O_3)		
A (0)	B (1)	C (2)
239	5	0
5	281	1
2	3	415

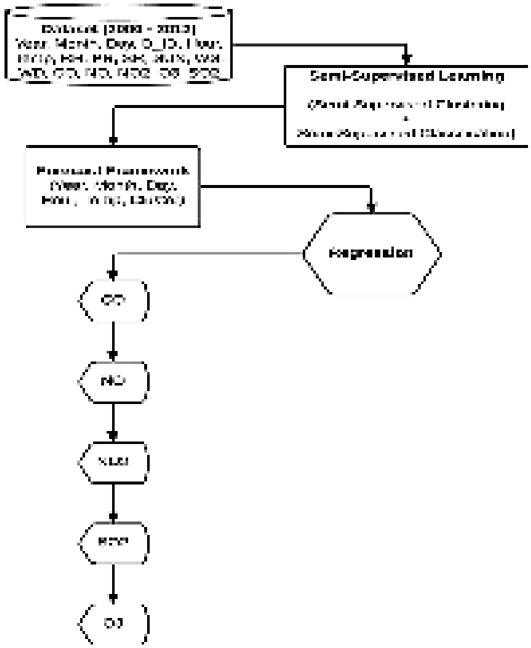

Fig. 2. Graph of the developed algorithm

The main advantage of our approach is that it runs only for one measuring station aiming to offer an overall classification for the whole area under study, much faster and in a simpler way. The same method can be applied for any other station. Also the Semi-Supervised Learning employed runs effectively by using only three classes whereas the fuzzy c-means required 5 classes and the SOM needed 9 classes in order to determine effectively the extreme pollutants' groups. The hypothesis that a pollution record is either harmless or dangerous for the public health, being related to high concentrations of primary or secondary pollutants is rather rational, flexible and moreover effective. The following Tables 6 and 7 present a comparison between the performance of the herein proposed method SSC^2-HAQS and the SOM, GAS, FUZZY and Unsupervised SOM that were applied in a previous research effort of our team [3]. The SSC^2-HAQS has better performance (for 3 out of 5 features). Specifically, it is more reliable for the NO_2, O_3 and SO_2 whereas it is equally reliable (though a little worse) for the CO and NO cases. However, it is a good compromise since it is much faster it models the whole area with the use of a single measuring station and it requires fewer classes in order to group the extreme values effectively.

Table 4. Training results

Training (2000–2012) 44601 instances	R^2	RMSE
CO	0.82	0.81
NO	0.78	55.5
NO_2	0.84	12.1
O_3	0.91	10.07
SO_2	0.75	5.38

Table 5. Testing results

Testing (2013) 5098 instances	R^2	RMSE
CO	0.78	0.59
NO	0.82	37.34
NO_2	0.53	12.88
O_3	0.70	19.94
SO_2	0.12	3.35

The following table presents the number of data vectors assigned the extreme tag. The four approaches of our previous research [3] have used data from four measuring stations and thus they had one more feature (All Stations). The SSC^2-HAQS incorporates more data vectors in the extreme cluster except for the UNSUPERSOM, which has very bad performance according to the previous Tables 6, 7 and 8.

Table 6. Comparison of performance for the extreme datasets (Training)

Training comparison (2000–2012)	CO		NO		NO_2		O_3		SO_2	
	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
SOM	0.86	0.75	0.92	36	0.74	19.2	0.86	14	0.71	15.7
GAS	0.90	0.7	0.94	33	0.74	17.6	0.83	17.5	0.62	13.7
FUZZY	0.88	0.62	0.92	30.27	0.72	15.4	0.83	15.4	0.64	10.7
UNSUPER SOM	0.42	1.29	0.37	76.39	0.54	23.63	0.9	10.27	0.34	16.23
SEMI	0.82	0.81	0.78	55.5	0.84	12.1	0.91	10.07	0.75	5.38

5 Discussion–Conclusions

This work presents an innovative and effective method of analyzing high concentrations of air pollutants with a combined hybrid Semi-Supervised Learning system. The proposed approach was tested successfully, in classifying and also in forecasting the extreme primary and secondary pollutant values for the center of Athens. It uses a

Table 7. Comparison of performance for the extreme datasets (Testing)

Testing comparison (2013)	CO		NO		NO_2		O_3		SO_2	
	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
SOM	0.77	0.53	0.83	40	0.48	17.9	0.71	33.3	0.13	6.88
GAS	0.76	0.62	0.9	30.1	0.49	16.2	0.4	36.9	0.14	6.69
FUZZY	0.76	0.57	0.85	40.6	0.53	14.5	0.69	19.5	0.1	6.51
UNSUPER SOM	0.19	0.98	0.38	58	0.25	25.1	0.27	35.4	0.03	7.13
SEMI	0.78	0.59	0.82	37.34	0.53	12.88	0.7	19.94	0.12	3.35

Table 8. Comparison of the extreme records' number (number of records)

Number of extreme records	Training (2000–2012)		Testing (2013)	
	All stations	Athinas	All stations	Athinas
SOM	30077	3383	14129	4378
GAS	53589	9354	13965	4343
FUZZY	91440	24834	14273	3987
UNSUPER SOM	213058	51304	19950	7757
SEMI	-	44601	-	5098

sophisticated technique of combined learning, which ensures fast, robust and effective forecasting and classification performance. Moreover, it is a general model which does not require specific characteristics of the area under study. All the above, add generalization ability to the methodology which is easily adjustable and applicable to other areas (cities) of research. The SSC^2-HAQS employs a Semi-Supervised Learning algorithm which is considered a realistic machine learning method that can model the most serious problems of the real world, based on the essential peculiarities that might characterize them. A main innovation introduced by the proposed scheme, concerns the data classification in homogeneous classes (distinction between primary and secondary pollutants). This process is done based on a sample of few pre-classified data vectors, something that incorporates the hidden knowledge and the correlations between the features. This hybrid system was tested effectively, with data that have specific particularities as they originate from a period of financial crisis for Greece, which has a significant effect on air quality in major urban centers.

Future work could involve testing of the system data in other urban centers with different climatic conditions and moreover it should consider climate change scenarios in these regions. Additionally, it would be very important to apply a new weights learning algorithm which will modify and adjust them based on specificity rates that are deemed necessary for the local climate. Thus the system could be made more flexible in achieving results in future evaluations and investigations of a region.

References

1. Bougoudis, I., Iliadis, L., Papaleonidas, A.: Fuzzy inference ANN ensembles for air pollutants modeling in a major urban area: the case of Athens. In: Mladenov, V., Jayne, C., Iliadis, L. (eds.) EANN 2014. CCIS, vol. 459, pp. 1–14. Springer, Heidelberg (2014)
2. Bougoudis, I., Iliadis, L., Spartalis, S.: Comparison of self organizing maps clustering with supervised classification for air pollution data sets. In: Iliadis, L., Maglogiannis, L., Papadopoulos, H. (eds.) AIAI 2014. IFIP AICT, vol. 436, pp. 424–435. Springer, Heidelberg (2014)
3. Bougoudis, I., Demertzis, K., Iliadis, L.: Fast and low cost prediction of extreme air pollution values with hybrid unsupervised learning. Integr. Comput.-Aided Eng. **23**(2), 115–127 (2016). doi:10.3233/ICA-150505. IOS Press
4. Bougoudis, I., Demertzis, K., Iliadis, L.: HISYCOL a hybrid computational intelligence system for combined machine learning: the case of air pollution modeling in Athens. EANN Neural Comput. Appl. 1–16 (2016). doi:10.1007/s00521-015-1927-7
5. Roy, S.: Prediction of particulate matter concentrations using artificial neural network. Resour. Environ. **2**(2), 30–36 (2012). doi:10.5923/j.re.20120202.05
6. Robles, L.A., Ortega, J.C., Fu, J.S., Reed, G.D., Chow, J.C., Watson, J.G., Moncada-Herrera, J.A.: A hybrid ARIMA and artificial neural networks model to forecast particulate matter in urban areas: the case of Temuco, Chile. Atmos. Environ. **42**(35), 8331–8340 (2008)
7. Ordieres Meré, J.B., Vergara González, E.P., Capuz, R.S., Salaza, R.E.: Neural network prediction model for fine particulate matter (PM). Environ. Model Softw. **20**, 547–559 (2005)
8. Wahab, A., Al-Alawi, S.M.: Assessment and prediction of tropospheric ozone concentration levels using artificial neural networks. Environ. Model. **17**, 219–228 (2002)
9. Paschalidou, A., Iliadis, L., Kassomenos, P., Bezirtzoglou, C.: Neural modeling of the tropospheric ozone concentrations in an urban site. In: Proceedings of 10th International Conference Engineering Applications of Neural Networks, pp. 436–445 (2007)
10. Ozcan, H.K., Bilgili, E., Sahin, U., Bayat, C.: Modeling of trophospheric ozone concentrations using genetically trained multi-level cellular neural networks. Advances in Atmospheric Sciences, vol. 24, pp. 907–914. Springer, Heidelberg (2007)
11. Ozdemir, H., Demir, G., Altay, G., Albayrak, S., Bayat, C.: Prediction of tropospheric ozone concentration by employing artificial neural networks. Environ. Eng. Sci. **25**(9), 1249–1254 (2008)
12. Inal, F.: Artificial neural network prediction of tropospheric ozone concentrations in Istanbul, Turkey. CLEAN – Soil Air Water **38**(10), 897–908 (2010)
13. Paoli, C.: A neural network model forecasting for prediction of hourly ozone concentration in Corsica. In: Proceedings IEEE of 10th International Conference on EEEIC (2011)
14. Kadri, C., Tian, F., Zhang, L., Dang, L., Li, G.: Neural network ensembles for online gas concentration estimation using an electronic nose. IJCS **10**(2), 1 (2013)
15. Vong, C.-M., Ip, W.-F., Wong, P.-K., Yang, J.-Y.: Short-term prediction of air pollution in Macau using support vector machines. J. Control Sci. Eng. **2012**, 4 (2012). Article ID 518032
16. Xiao, F., Li, Q., Zhu, Y., Hou, J., Jin, L., Wang, J.: Artificial neural networks forecasting of PM 2.5 pollution using air mass trajectory based geographic model and wavelet transformation. Atmos. Environ. **107**, 118–128 (2015). doi:10.1016/j.atmosenv.2015.02.030. Elsevier
17. Zabkar, R., Cemas, D.: Ground-level ozone forecast based on ML. AIR040051 (2004)

18. Lopez-Rubio, E., Palomo, E.J., Dominguez, E.: Bregman divergences for growing hierarchical self-organizing networks. Int. J. Neural Syst. **24**, 4 (2014). 1450016
19. Menendez, H., Barrero, D.F., Camacho, D.: A genetic graph-based approach to the partitional clustering. Int. J. Neural Syst. **24**, 3 (2014). 1430008
20. Donos, C., Duemoelmann, M., Schulze-Bonhage, A.: Early seizure detection algorithm based on intractable EEG and random forest classification. IJNS **25**, 5 (2015). 1550023
21. Quirós, P., Alonso, P., Díaz, I., Montes, S.: On the use of fuzzy partitions to protect data. Integr. Comput.-Aided Eng. **21**(4), 355–366 (2014)
22. Sen, P., Namata, G., Bilgic, M., Getoor, L., Galligher, B., Eliassi-Rad, T.: Collective classification in network data. Assoc. Adv. AI **29**(3), 93 (2015)
23. Driessens, K., Reutemann, P., Pfahringer, B., Leschi, C.: Using weighted nearest neighbor to benefit from unlabeled data. In: Ng, W.-K., Kitsuregawa, M., Li, J., Chang, K. (eds.) PAKDD 2006. LNCS (LNAI), vol. 3918, pp. 60–69. Springer, Heidelberg (2006)