Neural Comput & Applic (2018) 29:375–388
https://doi.org/10.1007/s00521-017-3125-2

S.I. : EANN 2016

FuSSFFra, a fuzzy semi-supervised forecasting framework: the case of the air pollution in Athens

Ilias Bougoudis[1] · Konstantinos Demertzis[2] · Lazaros Iliadis[3] · Vardis-Dimitris Anezakis[2] · Antonios Papaleonidas[2]

Received: 15 February 2017 / Accepted: 18 June 2017 / Published online: 24 June 2017

Abstract Mining hidden knowledge from available datasets is an extremely time-consuming and demanding process, especially in our era with the vast volume of high-complexity data. Additionally, validation of results requires the adoption of appropriate multifactor criteria, exhaustive testing and advanced error measurement techniques. This paper proposes a novel Hybrid Fuzzy Semi-Supervised Forecasting Framework. It combines fuzzy logic, semi-supervised clustering and semi-supervised classification in order to model Big Data sets in a faster, simpler and more essential manner. Its advantages are clearly shown and discussed in the paper. It uses as few pre-classified data as possible while providing a simple method of safe process validation. This innovative approach is applied herein to effectively model the air quality of Athens city. More specifically, it manages to forecast extreme air pollutants' values and to explore the parameters that affect their concentration. Also it builds a correlation between pollution and general climatic conditions. Overall, it correlates the built model with the malfunctions caused to the city life by this serious environmental problem.

Keywords Air quality · Air pollution · Fuzzy logic · Semi-supervised learning · Semi-supervised clustering · Semi-supervised classification

1 Introduction

1.1 Effect of air pollution to the climate change

Apart from the natural processes, a main cause of climate change is pollution of the atmosphere by human activities that contribute to the increase in the greenhouse gases concentration. Therefore, the radiation increases and heat is trapped in the atmosphere, resulting in the enhancing of the natural greenhouse effect.

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. Human activities are related to the excessive use of fossil resources such as coal, lignite, oil and natural gas, the combustion of which releases vast amounts of CO_2 into the atmosphere. Overall transport is responsible for approximately 50% of the total emissions followed by industries and power plants [1]. Totally, 40% of the pollutants' concentration is related to CO_2, while the remaining 10% consists of other gases, mainly O_3 and CO. Moreover, the continuous deforestation contributes to the increase in greenhouse gases by 15% [1]. The production

✉ Konstantinos Demertzis
kdemertz@fmenr.duth.gr

Ilias Bougoudis
ibougoudis@iup.physik.uni-bremen.de

Lazaros Iliadis
liliadis@civil.duth.gr

Vardis-Dimitris Anezakis
danezaki@fmenr.duth.gr

Antonios Papaleonidas
apapaleo@fmenr.duth.gr

[1] Institute of Environmental Physics, DOAS Group, University of Bremen, Otto-Hahn-Allee 1, 28359 Bremen, Germany

[2] Lab of Forest Informatics, Democritus University of Thrace, 193 Pandazidou st., 68200 Orestiada, Greece

[3] School of Engineering, Department of Civil Engineering, Democritus University of Thrace, University Campus, Kimmeria, 67100 Xanthi, Greece

and use of synthetic chemical substances has disturbed irreparably the balance in the CO_2 cycle. These effects alter the natural protective shield covering the Earth, thereby retaining more and more energy which, in turn, increases the global average temperature, while the speed at which this increase occurs is substantially greater than any natural process. The result is the inability of natural systems to adapt to the new circumstances.

Air pollution in urban and industrial areas in Greece is a major environmental problem as it is associated mainly with population growth in cities, uncontrolled urban expansion and growth trends in industrial production.

The problem of air pollution in the Attica basin and particularly in the urban center of Athens–Piraeus is directly related to overpopulation, to industrial-craft activities and to the unfavorable topography. The yellow–brown photochemical cloud of Athens comprises smog, CO, SO_2, NO_x PM_{10}, $PM_{2.5}$ and O_3. Financial crisis is an additional air charge reason during the last 7 years, and it is related to the use of alternative heating methods, such as fireplaces and pellet burners.

An assessment of the effects of atmospheric pollution and also a thorough and comprehensive investigation that would lead to the forecasting of the extreme pollutants' values requires the analysis of the conditions that favor high concentrations, based on rational approaches. This research effort proposes an innovative analytical soft computing model toward estimation of the pollutants concentrations. More specifically, it manages to forecast extreme air pollutants' values only with the usage of temporal and meteorological parameters as inputs. As a result, meaningful information can be extracted for the conditions under which these extreme pollution cases occur. The proposed approach will apply the FuSSFFra model for this purpose. It combines fuzzy logic, semi-supervised clustering and semi-supervised classification in order to label efficiently Big Data with a minimum pre-labeled dataset, predict extreme values of pollutants without any other pollutant as input and evaluate the labeling procedure in a novel semi-supervised way. The biggest advantage of this research effort though is that the proposed algorithm as a whole can be applied to any case as an evaluation method for unsupervised learning, independently of the environmental problem that it is implemented here to deal with.

1.2 Semi-supervised learning

The main disadvantage of the classical supervised machine learning methods is that they presuppose the existence of a large number of *pre-classified* data to properly train a model with adequate accuracy. Building the training-validation sets manually is a tedious and time-consuming process, especially intangible when we are dealing with a huge dataset. On the other hand, building a model by employing semi-supervised learning (SSL) involves training with a minimum number of pre-classified data records. This approach can potentially be quite more rational and suitable for many timely cases. This method focuses on the classification of the distribution of unlabeled data and on the parallel correction of misclassification errors based in already known available classes. The unlabeled data provide useful information for the exploration of the overall dataset, whereas the classified ones contribute to the learning process. Moreover, taking into account the substantial particularities of the most real-world problems, the success of learning with partial supervision depends on some basic assumptions imposed by each case, which can be modeled by proper SSL algorithms. This fact adds substantial merit and rationality to the SSL methods.

2 Related research

In an earlier research of our team [2], we have made an effort to get a clear and comprehensive view of air quality in the wider urban center of Athens and also in the Attica basin. This study was based on data that were selected from nine air pollution-measuring stations of the area during the temporal periods (2000–2004, 2005–2008 and 2009–2012). This method was based on the development of 117 partial ANNs, while their 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 pollutant. The results showed that this approach outperforms the other five ensemble methods.

Also, in previous work [3], we used one unsupervised learning method (self-organizing maps), in order to cluster our pollutants in groups. Our ultimate goal was to find the most isolated group, where we hoped that all the extreme values of pollutants were gathered. This specific group is the most important, as it contains vital information about the hazardous pollutants: the meteorological and chronological conditions under which the extreme pollutants occur. Moreover, we tried to evaluate the clustering above, using pattern recognition. In order to produce the groups above, we only used the pollutants as inputs. In the classification, we used every input we had for each record: 5 chronological inputs, 7 meteorological and 5 pollutants. We used classification because we wanted to see whether the clustering above is suitable for future use.

In addition, in [4], the EHF innovating forecasting system, which allows the prediction of extreme air pollutant values, was introduced and tested with real data records. Its main advantage is that though it takes no pollutants as inputs it manages to operate quite efficiently. Moreover, it uses a small number of inputs (7), which comprises 4 temporal inputs, air temperature, a station identification code (was determined automatically by geolocation-based services) and a cluster identification code. In order to produce the EHF model, we have used four unsupervised learning algorithms namely: SOM, neural gas ANN, fuzzy C-means and a fully unsupervised SOM algorithm. For every algorithm, we have searched for the most extreme cluster, which contained the most hazardous pollutant values EXPV. Thereafter, we 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. Bougoudis et al. [5] proposed 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.

Bougoudis et al. [6] proposed an innovative hybrid system of combined machine learning algorithms. They presented an ensemble system using combination of machine learning algorithms capable of forecasting the values of air pollutants. This approach improved the accuracy of existing forecasting models by using unsupervised machine learning to cluster the data vectors and trace hidden knowledge.

In some machine learning applications, using soft labels is more useful and informative than crisp labels. Soft labels indicate the degree of membership of the training data to the given classes. Often, only a small number of labeled data are available while unlabeled data are abundant. Therefore, it is important to make use of unlabeled data. Semi-supervised learning makes use of both labeled and unlabeled samples. Semi-supervised learning addresses the classification problem by using large amount of unlabeled data, together with the labeled data, to build better classifiers. It tries to overcome the need for a large labeled training set to learn accurate classifiers.

Krithara et al. [7] investigated a new extension of the probabilistic latent semantic analysis (PLSA) model for text classification where the training set was partially labeled. The proposed approach iteratively labels the unlabeled documents and estimates the probabilities of its labeling errors. These probabilities are then taken into account in the estimation of the new model parameters before the next round.

Ashfag et al. [8] proposed a novel fuzziness-based semi-supervised learning approach by utilizing unlabeled samples assisted with supervised learning algorithm to improve the classifier's performance for the IDSs. A single hidden layer feedforward neural network (SLFN) was trained to output a fuzzy membership vector, and the sample categorization (low, mid and high fuzziness categories) on unlabeled samples was performed using the fuzzy quantity.

Yan and Chen [9] proposed the use of labeled data, and the exploration of the constraints generated from the labels during the clustering process. They formulated the clustering process as a constrained optimization problem and proposed a novel semi-supervised fuzzy co-clustering algorithm which was incorporated with a few category labels to handle large overlapping text corpus. Zheng and Luo [10] presented a novel semi-supervised fuzzy SVM clustering framework where the spatial distribution information of the unlabeled samples and the prompted information of the labeled samples were integrated to obtain better results. Le et al. [11] proposed a semi-supervised learning method, fuzzy entropy semi-supervised SVDD (FS3VDD), to extend SVDD to cope with partially labeled datasets. The learning model employed fuzzy membership and fuzzy entropy to help the labeling of the unlabeled data. Yan et al. [12] presented a new method, called semi-supervised fuzzy relational classifier, which combined semi-supervised clustering and classification together. In the proposed semi-supervised fuzzy relational classifier, they employed the semi-supervised pair wise-constrained competitive agglomeration (PCCA) to replace FCM to obtain clusters fitting the user expectations without specifying the exact cluster number. In addition, they incorporated the fuzzy class labels of unlabeled data into the classification mechanism to improve its performance.

Benbrahim [13] proposed a fuzzy semi-supervised support vector machines (FSS-SVM) algorithm. It tried to overcome the need for a large labeled training set to learn accurate classifiers. For this, it used both labeled and unlabeled data for training. It also modulated the effect of the unlabeled data in the learning process. Empirical evaluations showed that by additionally using unlabeled data, FSS-SVM required less labeled training data than its supervised version, support vector machines to achieve the same level of classification performance. Also, the incorporated fuzzy membership values of the unlabeled training patterns in the learning process had positively influenced the classification performance in comparison with its crisp variant.

El-Zahnar and El-Gayar [14] proposed an approach for Fuzzy-Input Fuzzy-Output classification in which the classifier can learn with soft-labeled data and can also produce degree of belongingness to classes as an output for each pattern. They investigated two semi-supervised multiple classifier frameworks for this classification purpose. Results showed that semi-supervised multiple classifiers can improve the performance of fuzzy classification by making use of the unlabeled data. Finally, Jamalabadi et al. [15] proposed a new reasoning method for fuzzy classifiers referred to as competitive interaction reasoning (CIR), that employs the cumulative information provided by all fuzzy rules and adjusts the decision boundaries as if the membership functions are directly modified. The proposed CIR significantly improves classification accuracy without compromising interpretability of the fuzzy classifier.

Cordeiro et al. [16] presented a new semi-supervised segmentation algorithm based on the modification of the GrowCut algorithm to perform automatic mammographic image segmentation once a region of interest is selected by a specialist. They used fuzzy Gaussian membership functions to modify the evolution rule of the original GrowCut algorithm, in order to estimate the uncertainty of a pixel being object or background. Yan et al. [17] proposed a new semi-supervised fuzzy kernel clustering algorithm (SFKC) based on some modifications of the fuzzy clustering methods. In Tanaka et al. [18], the virtual sample approach was adopted in semi-supervised fuzzy co-clustering. The goal was to reveal object-item pairwise cluster structures from co-occurrence information among them. Honda et al. [19] proposed a novel framework for performing fuzzy co-clustering of co-occurrence information with partial supervision, which was induced by multinomial mixture concept. Jensen et al. [20] proposed a novel approach for semi-supervised fuzzy-rough feature selection where the object labels in the data may only be partially present. The approach also has the appealing property that any generated subsets are also valid when the whole dataset is labeled. Le et al. [21] showed how to apply the fuzzy theory to the proposed semi-supervised one-class classification method for efficiently handling noises and outliers. Diaz-Valenzuela et al. [22] introduced fuzzy HSS, a semi-supervised hierarchical clustering approach that uses fuzzy instance-level constraints. These constraints are external information on the shape of fuzzy must-link and fuzzy cannot-link restrictions. They allow uncertainty when indicating whether two instances of a dataset belong to the same group.

Bchir et al. [23] proposed a novel method of learning nonlinear distance functions with side information while clustering the data. The proposed algorithm learns the underlying cluster-dependent dissimilarity measure while finding compact clusters in the given dataset.

3 Theoretical background

FuSSFFra applies a hybrid model, which employs well-established algorithms, optimally combined in order to create a faster and more flexible integrated fuzzy semi-supervised learning system. The most important innovation and advantage of the proposed approach is the easy validation of the classification process for a first time seen dataset, based on robust measurable factors. The theoretical background of the system's core is presented in the next paragraphs.

3.1 Naive Bayes classifiers

The Naive Bayes classifier is a practical learning method based on a probabilistic representation of a data structure, representing a set of random variables and their hypothetical independence, in which complete and combined probability distributions are substantiated. The objective of the algorithm is to classify a sample X in one of the given categories C_1,C_2,…,C_n using a probability model defined according to the theory of Bayes. These classifiers make probability assessment rather than forecasting, which is often more useful and effective. Here the projections have a score, and the purpose is the minimization of the expected cost. Each category is represented by a prior probability. We make the assumption that each sample X belongs to a class $\mathrm{C_i}$ and based on the Bayes theory we estimate the posteriori probability. The quantity P describing a Naive Bayes classifier for a set of samples expresses the probability that c is the value of the dependent variable C, based on the prices $x = (x_1, x_2,\ldots, x_n)$ of the properties $X = (X_1, X_2,\ldots, X_n)$, and it is given by relation (1) where the characteristics x_i are considered as independent:

$$P(c|x) = P(c) \cdot \prod_{i}^{n} P(x_i|c)(1) \tag{1}$$

The estimation of the above quantity for a set of N examples is done by using relations 2, 3 and 4:

$$P(c) = \frac{\mathrm{N}(c)}{\mathrm{N}} \tag{2}$$

$$P(x_i|c) = \frac{\mathrm{N}(x_i, c)}{\mathrm{N}(c)} \tag{3}$$

For a characteristic x_i with discrete values, the probability is estimated by Eq. 4.

$$P(x_i|c) = g(x_i, \mu c, \sigma c2) \tag{4}$$

where $N(c)$ is the number of examples that have the value c for the depended variable, $N(x_i,c)$ is the number of cases that have the values x_i and c for the characteristic $\mathrm{X_i}$ and the depended parameter, respectively, and $g(x_i, \mu c, \sigma c2)$ is the Gaussian probability density function with an average value μc and variance σc for the characteristic x_i.

3.2 Collective classification

Collective classification [24] is a combinatorial optimization problem, in which we are given a set of nodes, $V = \{V_1,\ldots,V_n\}$ and a neighborhood function N, where

$Ni \subseteq V \backslash \{Vi\}$. 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. Our task is to label the nodes $Y_i \in Y$ with one of a small number of labels, $L = \{L_1,\ldots,L_q\}$; we will use the shorthand y_i to denote the label of node Y_i.

3.3 Fuzzy clustering

According to Zadeh [25–27], every element "x" of the Universe of discourse "X" belongs to a fuzzy set (FS) with a degree of membership in the closed interval [0,1]. Thus, function 5 is the mathematical foundation of a FS.

$$S = \{(x, \mu s(x)/\mu s\{[0,1] : x\}\mu s(x)\}. \tag{5}$$

Function 6 is an example of a typical Triangular Fuzzy Membership Faction (FMF). It must be clarified that the "a" and "b" parameters have the values of the lower and upper bounds of the raw data, respectively.

$$\mu_s(X) = \begin{cases} 0 \quad \text{if } X < \alpha \\ (X-a)/(c-a) & \text{if } X \in [a,c) \\ (b-X)/(b-c) & \text{if } X \in [c,b) \\ 0 \quad \text{if } X > b \end{cases} \tag{6}$$

According to the typical (crisp) classification methods, each sample can be assigned only to one class. Thus, the class membership value is either 1 or 0. In general, classification methods reduce the dimensionality of a complex dataset by grouping the data into a set of classes.

In fuzzy classification, a sample point can be assigned to many classes with a different degree of membership.

The fuzzy c-means clustering algorithm initially gives random values to the cluster centers, and then it assigns all of the data points to all of the clusters with varying degrees of membership (DOM) by measuring the Euclidean distance.

The Euclidean distance of each data point x_i from the center of each cluster $c_1\ldots c_j$ *is* calculated based on Eq. 7 [28].

$$d_{ji} = \| x_i - c_j \|^2 \tag{7}$$

where d_{ji} is the distance of x_i from the center of the cluster c_j.

Then the DOM of each data point to each cluster is estimated based on Eq. 8

$$\mu_j(x_i) = \frac{\left(\frac{1}{d_{ji}}\right)^{\frac{1}{m-1}}}{\sum_{k=1}^{p}\left(\frac{1}{d_{ki}}\right)^{\frac{1}{m-1}}} \tag{8}$$

where m is the fuzzification parameter with values in the interval [1.25, 2] [28]. The values of m specify the degree of overlapping between the clusters. The default value of m is equal to 1.2. The algorithm has the following direct restriction in the DOM of each point [28]. See Eq. 9.

$$\sum_{j=1}^{p} \mu_j(x_i) = 1 \quad i = 1,2,3,\ldots k \tag{9}$$

where p is the number of the clusters, k is the number of the data points, x_i is the ith point and $\mu_j\ (x_i)$ is a function that returns the degree of membership of point x_i in the jth cluster $i = 1,2,\ldots k$.

Then the centers are estimated again. Eq. 10 is used for the re-estimation of the values of new cluster centers [28].

$$c_j = \frac{\sum_i \left[\mu_j(x_i)\right]^m x_i}{\Sigma_i \left[\mu_j(x_i)\right]^m} \tag{10}$$

where c_j is the center of the jth cluster with ($j = 1,2\ldots p$), and x_i is the ith point [28]. This is an iterative algorithm, and the whole process is repeated till the centers are stabilized.

4 The proposed system

4.1 The FuSSFFra algorithm

The FuSSFFra is an innovative hybrid algorithm based on the combination of soft computing approaches. It is described herein for the first time in the literature.

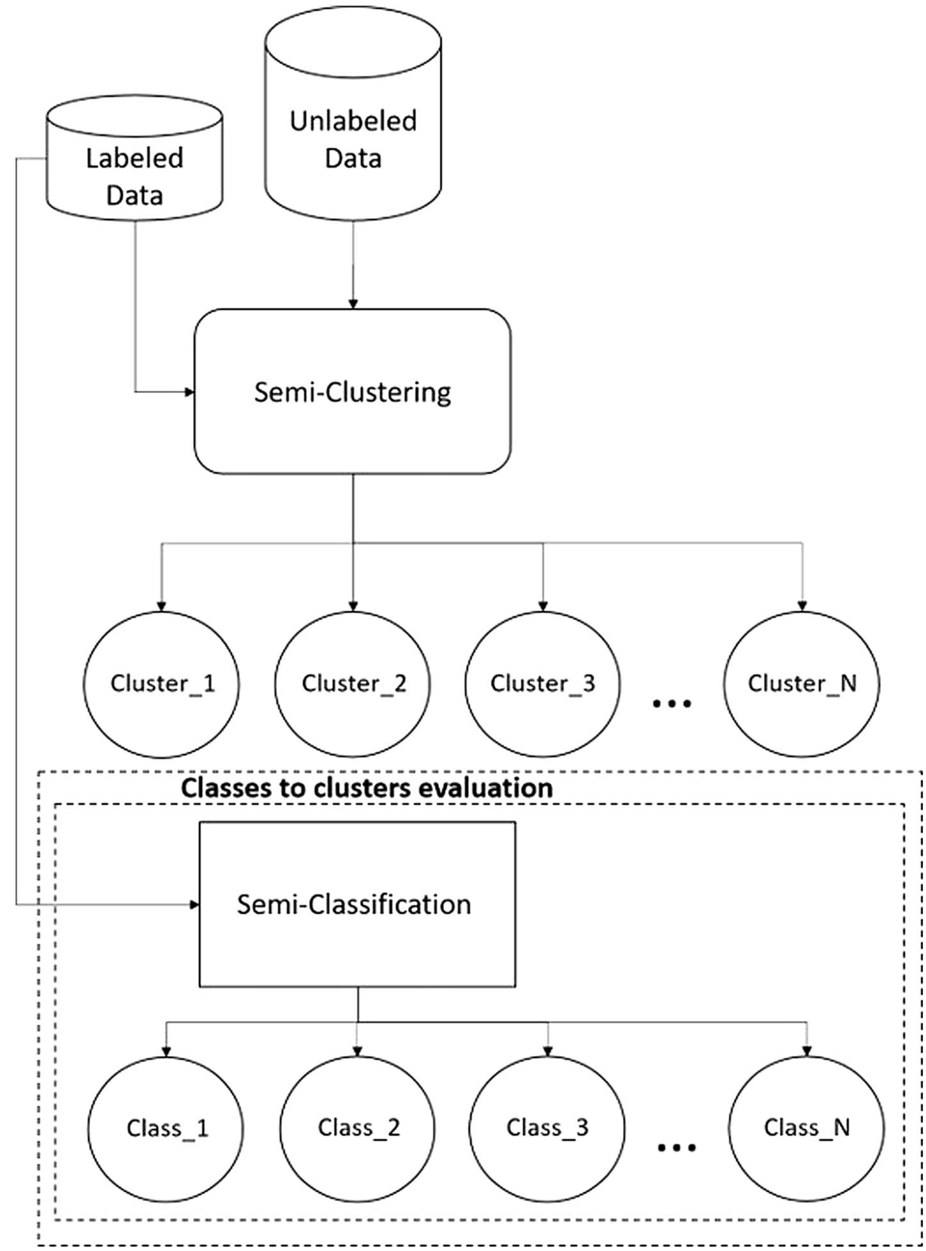
Let us consider a supervised learning case with a training set of size $N\ \{X,Y\} = \{x_i, y_i\}_{i=1}^{N}$, where $x_i \in R^{n_i}$ and y_i is a binary vector of size n_o. It must be clarified that *i and* n_o are the dimensions of the input and output, respectively.

The FuSSFFra initially performs semi-supervised clustering (SSC). This means that cluster assignments may be already known for some subset of the data. The final aim is the classification of the unlabeled observations to the appropriate clusters, using the known assignments for this subset of the data. At the same time, the algorithm produces the degree of membership of each record to its cluster.

The clustering validation process is performed by employing the "*classes to clusters*" (CL_A_U) method that adopts SSC. Originally, a minimum data sample is used comprising of the clusters derived from the SSC process (labeled data). The remaining unlabeled data are used to dynamically form and adjust the classes based on their DOM.

Actually, the CL_A_U approach assigns classes to the clusters, based on the majority value of the class attribute within each cluster. The class attribute is treated like any other attribute, and it is a part of the input to the clustering algorithm.

Fig. 1 The proposed FuSSFFra



The objective is to determine whether the selected clusters match the specified class data. In the CL_A_U evaluation, the system is informed on which attribute is a predetermined "class." Then this is removed from the data before passing to the SSC algorithm. The CL_A_U evaluation finds the minimum error of mapping classes to clusters (where only the class labels that correspond to the instances in a cluster are considered) with the constraint that a class can only be mapped to one cluster. Figures 1 and 2 presents an overall description of the FuSSFFra system.

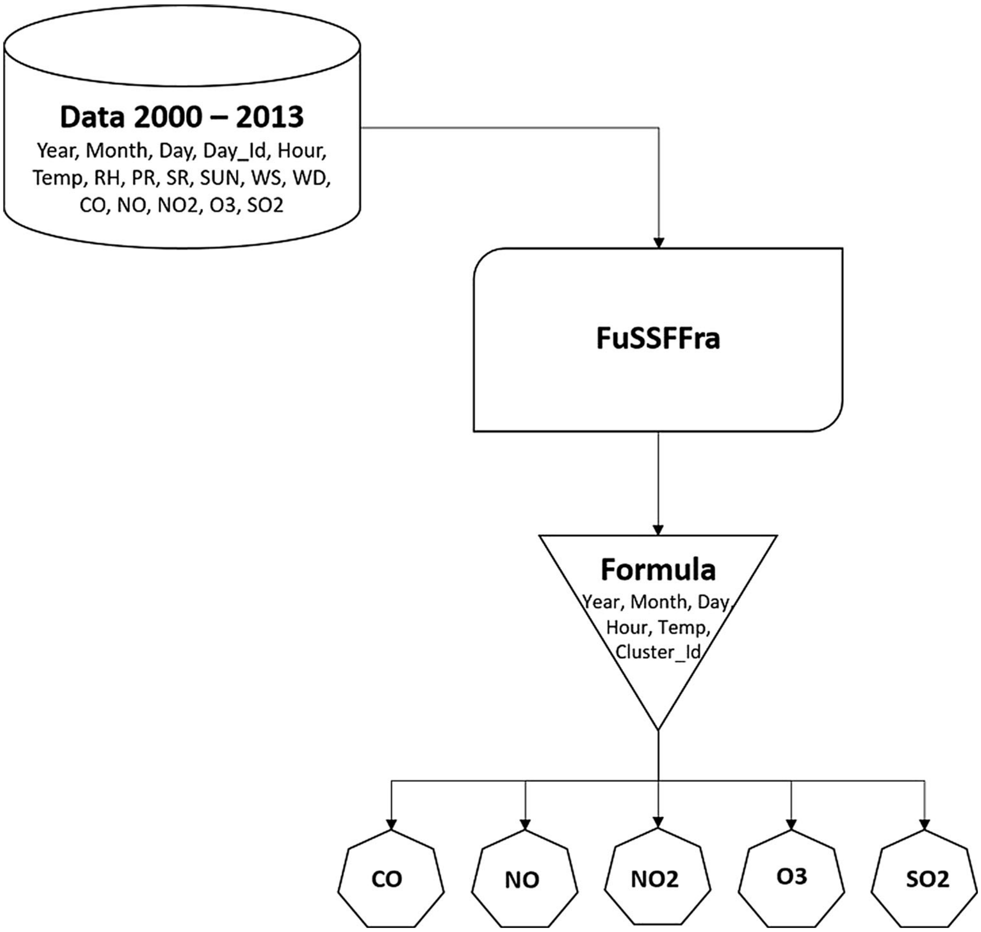
The 10-fold cross-validation (10_FCV) is employed in this stage in order to obtain performance indices. The idea behind k_FCV (where k is a positive integer) is to create a number of k partitions (folds) of the sample observations. Then the model is trained on k–1-folds, and the test error is predicted on the left-out partition (LOP). The process is repeated k times (after interchanging the LOP), and the result is averaged. The process is referred to as leave-one-out cross-validation. Taking the average of the *k* accuracy scores is a macro-average.

The emerged classes are fuzzified by assigning them proper Linguistics, in order to obtain a realistic coherence between the associated values of the dataset under study.

The whole process is presented in details in Algorithm 1 below.

```
Algorithm 1. The FuSSFFra Algorithm
Inputs: Input labeled data $D_l$, clusters of the labeled data $L_l$ and a set of unlabeled data $D_u$
    Step 1: % Initialization of clusters
        Identify the discrete number of clusters based on $L_l$
        For every cluster, create matrices with the mean and standard deviation of all $D_l$
    Step 2: % Calculate the new centers of the clusters
        For every cluster, recreate these matrices, based on the testing data $D_u$
        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 $D_u$
        Sum all these variables for each cluster
    Step 3: % Calculate the winner cluster for each record
        For every testing data $D_u$, find the minimum value of the summary calculated before.
        % Calculate the fuzzy membership values for every cluster for every record
        For every testing data $D_u$ and for every class, divide the mean matrix with the sum of the
            values calculated before (normalization probability – membership value)
Outputs: Winner cluster for each testing data $D_u$, $C_u$ and fuzzy membership values for every cluster
        for every testing data $D_u$, $F\_M\_V_{u,j}$ (j the number of clusters)
    Step 5: % Validation of the clustering process
        Repeat Steps 1 – 3 from the previous part, only this time from $D_u \rightarrow D_l$, using $C_u$ as labels
Output: Winner cluster for each testing data $D_l$, $L2_l$
    Step 6:
        For every initially labeled data $D_l$:
        Compare the initial label $L_l$ with $L2_l$
        Create confusion matrix based on these comparisons
    Step 7:
        Repeat Steps 5 - 6 for every $D_w$ of $D_u$
    % Generalization of the amount of the extreme cases, based on the fuzzy membership values
Inputs: The winner class for every record ($C_u$) and the fuzzy membership values for each record
        ($F\_M\_V_{u,j}$)
    Step 8:
        For every record:
        If $max(F\_M\_V_{u,j}) = A$ AND $F\_M\_V_{u,A} - max2(F\_M\_V_{u,j}) <= threshold$, then
        % $max2(F\_M\_V_{u,k}) = k$, the second biggest membership value
        Change the winner class for this record to k ($C_u = k$)
Outputs: Updated winner cluster for each record $C_u$
```

Fig. 2 Graphical representation of the Formula methodology


Data 2000 – 2013
Year, Month, Day, Day_Id, Hour, Temp, RH, PR, SR, SUN, WS, WD, CO, NO, NO2, O3, SO2
FuSSFFra
Formula
Year, Month, Day Hour, Temp, Cluster_Id
CO
NO
NO2
O3
SO2

5 The case of the air pollution in Athens

The FuSSFFra proposed herein offers a novel and suitable modeling approach for complex cases like air pollution of an urban center. It is important to examine the dependencies of this case, on the assumptions and characteristics of both the prevailing weather conditions and the overall timing in relation to the area concerned.

In fact, this research effort refers to the modeling and forecasting of extreme air pollutants' values for the city of Athens, which is a problem of high complexity [2–6, 29].

The data used are related to 14 years (2000–2013). They come from the "Athinas" station, which is located in the heart of the city, and it offers a typical city center air pollution image. More specifically, this station counts hourly measures of CO, NO, NO_2, O_3 and SO_2. Co is measured in mg/m^3, whereas for all other pollutants $\mu g/m^3$ is used.

The proposed semi-supervised modeling approach was used for the years 2000–2012, whereas the 2013 data were used for testing the efficiency of the developed model. Additionally, each record contains the following data: year, month, day, hour and temperature. Table 1 presents descriptive statistics of the 2000–2012 measurements.

Table 1 Statistical analysis for the period 2000–2012

2000–2012 (89364)	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	8	60	3	2
Count_mode	4941	2358	1508	6649	10451
Average	1.83	59.31	63.57	32.94	9.64
SD	1.48	90.06	27.11	28.67	9.39

Table 2 Statistical analysis for the year 2009

2012 (8728)	CO	NO	NO_2	O_3	SO_2
Max	10.4	660	323	174	56
Min	0.1	1	6	1	2
Mode	0.8	1	69	3	4
Count_mode	619	323	167	986	2173
Average	1.50	52.62	66.46	32.59	6.29
SD	1.06	80.39	25.10	26.94	3.53

Table 3 Statistical analysis for the year 2013

2013 (8146)	CO	NO	NO_2	O_3	SO_2
Max	10.2	539	118	151	39
Min	0.2	1	2	1	2
Mode	0.7	6	37	3	5
Count_mode	682	327	212	496	1668
Average	1.35	41.06	42.75	36.61	6.77
SD	0.96	62.54	16.18	26.71	2.93

Table 4 Statistical analysis of the extreme fuzzy semi-dataset 2000–2012 (Cluster 1)

2000–2012 (21635)	CO	NO	NO_2	O_3	SO_2
Max	21.4	908	377	139	259
Min	0.1	1	4	1	2
Mode	2.7	52	76	3	9
Count_mode	756	159	444	4908	979
Average	3.62	160.50	88.48	9.77	19.06
SD	1.911	133.81	28.77	12.38	13.53

Table 5 Statistical analysis of the extreme fuzzy semi-dataset 2000–2012 (Cluster 2)

2000–2012 (28003)	CO	NO	NO_2	O_3	SO_2
Max	3.1	33	153	253	35
Min	0.1	1	3	11	2
Mode	0.7	7	34	49	2
Count_mode	3357	2109	787	514	5942
Average	0.86	9.73	42.29	63.12	6.04
SD	0.38	5.28	15.35	25.68	4.40

Tables 2, 3, 4 and 5 have the same statistics for the years 2009 and 2013. The year 2009 was chosen for the determination of the initial classes because it contains the highest number of data records. The year 2013 as we have already mentioned was selected to be used for the testing process. The brackets in the first left square of all tables contain the number of the available data records.

It should be noted that after the procedure followed by the use of the FuSSFFra algorithm, the prediction was made by the Formula, *a heuristic variable subset of predictors*, which was proposed and presented in a previous research of our team [4].

5.1 Brief description of the Formula subset of predictors

Typical pollutants forecasting procedures require measurements at the level of independent variables, which are usually derived from monitoring stations or specialized hardware such as sensors.

In a previous research effort of our group [4], we have introduced a model that considers a special set of data named *Formula,* in order to offer to the society the capacity of being warned on extreme air pollution cases by using cheap hardware like mobile devices, without requiring real-time feedback of data.

The *Formula* is a particularly effective *variable subset of predictors* used to forecast extreme pollutants' values, in which the independent variables do not require continuous measurements from specialized hardware or software. The employment of Formula requires low resources, shorter training times and enhanced generalization by reducing chances for overfitting [4]. This subset comprises the following *independent variables*: *year, month, day, hour, airtemp and Cluster_Id*. In [4], the model was developed with *Formula* data for the years 2000–2012 and then it was tested to predict the values for *Formula* for 2013.

Indicatively, it should be mentioned that the value of R^2 was 0.76 in testing for CO, 0.90 for NO and 0.71 for O_3 by using the neural gas algorithm for the two first cases and the self-organizing maps for the third [4].

Herein, the proposed semi-fuzzy algorithm is employed to assign a *Cluster_Id* to each record of the available dataset. The *Cluster_Id* value is used to characterize each record as extreme or not, in terms of the primary or secondary air pollutants' values. After this successful clustering, the Formula dataset is formed and it can be used.

This is a very special and clearly useful innovation, which allows the free adoption and use of the *Formula* dataset by low-cost devices (e.g., smartphones). This subset of the initial available data (after a proper processing) can be used to forecast potential extreme values of CO, NO, NO_2, O_3, SO_2, without the input of any other pollutants' values coming from sensors. Figure 1 presents the methodology.

5.2 Athens air pollution modeling with the FuSSFFra

Once the grouping has been completed with the application of FuSSFFra, each record acquired an extra variable corresponding to its class, based on the pollutants' values. Totally, three Cluster_IDs were formed namely:

$Cluster\_Id = 0$ for **negligible** values of pollutants

$Cluster\_Id = 1$ for **extreme** values of the primary pollutants $(\mathrm{CO},\ \mathrm{NO},\ \mathrm{NO_2},\ \mathrm{SO_2})$

$Cluster\_Id = 2$ for **extreme** values of the secondary pollutants $(\mathrm{O_3})$

So, according to *steps 7 and 8* of the FuSSFFra algorithm, the records that were classified as "*negligible*" (when class 0 was prevalent) but their DOM to class 0 differed by their DOM to the next one (1 for primary or 2 for secondary) $\leq$0.2 were converted to extreme (value 1 or value 2, depending on the occasion). The 0.2 value was assigned based on the fact that the value 1 is the absolute degree to which a record can belong to a class. Thus, the 0.2 value is the highest 20% limit for this record to be a member of the second class. The same boundary has been used in similar cases in the literature several times [30, 31].

In this research, class A (*Cluster_Id = 0*) included the non-extreme cases, class B (*Cluster_Id = 1*) included the records with extreme values for the primary pollutants (CO, NO, NO_2, SO_2), and class C (*Cluster_Id = 2*) included the extreme values of the secondary pollutants (O_3). As a result, because we wanted our model to generalize based on the fuzzy algorithm it includes, we compared the difference between the membership value of each record to class A to its second highest membership value; if the subtraction result was <=0.2, then we changed the class of this record to 1 or 2, making it extreme. This resulted in the addition of 1406 extreme records in our dataset.

Thus, the records having Cluster_Id equal to 1 or 2 formed the extreme dataset, from which we extracted only the values corresponding to the *Formula* features namely: *year, month, day, hour, airtemp and Cluster_Id* (Demertzis et al. 2015). The *Formula* dataset is comprised of 49,638 records for the years 2000–2012 more than half of the initial ones. Then after creating a model for the period 2000–2012, we used it to forecast the extreme pollutants' values for 2013. We have employed the semi-supervised sub-clustering algorithm (SSSCA) to confirm the validity of clustering. The following tables support the results of this effort.

In the next step, we have used the *semi-supervised classification* algorithm (*classes to clusters evaluation*) to support the validity of this method. The algorithm works as follows: After running the semi-clustering sub-algorithm SSSCA which assigned clusters to all the records for all of the years, the semi-supervised sub-algorithm was used to estimate the class of each record on an annual basis. The input was all the records and their classes for the previous years.

After its execution, it compared its output with the ones of the semi-clustered sub-algorithm that was initially applied and its output for each year was not presented. Table 6 shows the comparison of the results between the two approaches.

Table 6 Total confusion matrix for the assignment of the classes

Year (instances)	Correct assignments (Percentage)	Incorrect assignments (Percentage)
2000 (5884)	5663 (96.2)	221 (3.75)
2001 (8053)	7763 (96.3)	290 (3.6)
2002 (6543)	6282 (96.0)	261 (3.9)
2003 (1238)	1158 (93.5)	80 (6.4)
2004 (3811)	3669 (96.2)	142 (3.7)
2005 (7787)	7504 (96.3)	283 (3.6)
2006 (7744)	7425 (95.8)	319 (4.1)
2007 (7036)	6770 (96.2)	266 (3.7)
2008 (8464)	8179 (96.6)	285 (3.6)
2009 (8728)	8605 (98.5)	123 (1.4)
2010 (6916)	6705 (96.9)	211 (3)
2011 (8473)	8096 (95.5)	377 (4.4)
2012 (8687)	8440 (97.1)	247 (2.8)
2013 (8146)	7912 (79.1)	234 (2.8)

Table 7 Overall confusion matrix for the assignment of the classes

Confusion matrix 97510 instances (0 normal values) (1 extreme primary) (2 extreme O_3)

A (0)	B (1)	C (2)
41356	426	1031
975	21466	5
901	1	31349

The following overall confusion matrix presents the effectiveness of the above approach for all 14 years. It is obvious that the number of correct classifications is very high and the method proves to be very promising Table 7.

6 Results and comparative analysis

6.1 Results

After the class assignment to each record, we tried to test the application of the *Formula* dataset, which was developed by our group in previous work [4], for predicting extreme pollutant values. The same features were used as described in the previous chapter. According to [4], these parameters should be enough to model efficiently the extreme air pollutants' values. Feedforward artificial neural networks (FFNNs) were employed for this purpose. A FFNN was used for each pollutant. The input is comprised of the parameters of the Formula set. The hidden neurons were ten in a single layer, whereas the *Tansig* transfer function was used with the *Trainlm* training function and the *Learngdm* learning one. Root-mean-square error (RMSE) was used as a metric for convergence. Table 8 presents the training results for each pollutant separately.

Additionally, another overall FFNN was used with the same input features and with 13 hidden neurons, in order to estimate the extreme values of all the pollutants; thus, it had five output neurons. In both cases, the ANN was trained for the period 2000–2012 and they were tested for 2013 without seeing the actual output for 2013. Table 9 presents the testing results.

6.2 Comparative analysis

From the above, we conclude that the combination of our fuzzy semi-algorithm with the employment of the *Formula* dataset gives satisfactory results toward the development of a *reliable, easy to use, fast and cheap low resource* predictive model, thus enhancing our effort to anticipate future extreme pollutant values for everyday use by people, without the presence of other pollutants as inputs.

Table 8 Training results

Training (2000–2012) 49638 instances	R^2	RMSE
CO	0.8	0.81
NO	0.79	52.40
NO_2	0.83	12.91
O_3	0.90	10.70
SO_2	0.77	5.37
ALL	0.73	33.47

Table 9 Testing results

Testing (2013) 5059 instances	R^2	RMSE
CO	0.78	0.54
NO	0.84	31.43
NO_2	0.50	13.50
O_3	0.69	15.48
SO_2	0.13	3.34
ALL	0.84	16.48

In this point, it was considered useful to compare the results of this effort with results obtained from previous similar research of our team [4, 5], where we have used other clustering algorithms which provided reliable results, but they did not offer a comprehensive validation approach. The difference with these approaches is that they used data from four measuring stations and thus they had one additional feature as input, namely the *station_id*).

The algorithm proposed herein offers an innovative approach compared to the previous ones (see [4]) as it considers only 10% of the data already assigned to classes in order to classify the rest of them. Also the method proposed in this research is more flexible compared to the one of [5] as it uses fuzzy logic in order to estimate not only the proper class for each record but to calculate its degree of membership. This offers a better generalization degree.

Tables 10 and 11 offer a comparison between this approach and the ones described in [4] and [5] for the estimation of the extreme values.

Looking at the above tables, we see that the fuzzy–semi-algorithm offers similar results with the algorithm that was used in [4]. However, herein the incorporation of fuzzy logic offers flexibility that enables better generalization capacity. Totally, 49,638 records were considered as extreme for the period 2000–2012 in a total of 89,364 ones. Tables 12, 13 and 14 shows the total number of extreme records for each algorithm for the city center and for the whole city of Athens (overall).

6.3 Fuzzy linguistics

The obtained results were fuzzified to correspond to proper Risk Linguistics (RL). More specifically, the predicted pollutants' values for the year 2013 were fuzzified in four RL namely: low, medium, high and extreme and it was checked whether the obtained RL was compatible with the

Table 10 Comparison between extreme datasets (training)

Training comparison (2000–2012)	CO		NO		NO_2		O_3		SO_2		ALL	
	R^2	RMSE	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	0.88	23
Gas	0.90	0.7	0.94	33	0.74	17.6	0.83	17.5	0.62	13.7	0.92	19.5
Fuzzy	0.88	0.62	0.92	30.27	0.72	15.4	0.83	15.4	0.64	10.7	0.92	17.12
Unsuper SOM	0.42	1.29	0.37	76.39	0.54	23.63	0.9	10.27	0.34	16.23	0.62	37.23
Semi	0.82	0.81	0.78	55.5	0.84	12.1	0.91	10.07	0.75	5.38	0.74	31.96
Fuzzy–semi	0.8	0.81	0.79	52.4	0.83	12.91	0.90	10.70	0.77	5.37	0.73	33.47

Table 11 Comparison between extreme datasets (testing)

Testing comparison (2013)	CO		NO		NO_2		O_3		SO_2		ALL	
	R^2	RMSE	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	0.86	25.5
Gas	0.76	0.62	0.9	30.1	0.49	16.2	0.4	36.9	0.14	6.69	0.66	20.1
Fuzzy	0.76	0.57	0.85	40.6	0.53	14.5	0.69	19.5	0.1	6.51	0.85	18.28
Unsuper som	0.19	0.98	0.38	58	0.25	25.1	0.27	35.4	0.03	7.13	0.45	34.7
Semi	0.78	0.59	0.82	37.34	0.53	12.88	0.7	19.94	0.12	3.35	0.82	22.99
Fuzzy–semi	0.78	0.54	0.84	31.43	0.50	13.5	0.69	15.48	0.13	3.34	0.84	16.48

Table 12 Comparison between extreme datasets (number of data records)

Number of extreme records	Training (2000–2012)		Testing (2013)	
	All Stations	Athinas station (city center)	All Stations	Athinas station (city center)
Som	30.077	3.383	14.129	4.378
Gas	53.589	9.354	13.965	4.343
Fuzzy	91.440	24.834	14.273	3.987
Unsuper som	213.058	51.304	19.950	7.757
Semi	–	44.601	–	5.098
Fuzzy–semi	–	49.638	–	5.059

Table 13 Linguistics boundaries for each air pollutant

Pollutants	Low	Medium	High	Extreme
CO	0.2802–1.883	0.681–2.284–3.886	2.684–4.287	>=4.2871
NO	5.178–109.5	31.27–135.6–239.9	161.7–266.1	>=266,372
NO_2	13.2–43.33	20.73–50.86–80.98	58.4–88.55	>=88.5517
O_3	1.92–50.61	14.09–62.78–111.5	74.96–123.7	>=128.605
SO_2	2.549–7.509	3.789–8.747–13.71	9.989–14.95	>=15.155

Table 14 Number of high and extreme linguistics (fuzzy sets)

2013	High	Extreme
CO	347	105
NO	331	112
NO_2	380	81
O_3	78	1
SO_2	187	61
ALL	2257	410

ones of the clusters. In this effort, four fuzzy sets were built to classify the pollutants' values using triangular membership functions (TFMF).

The characteristic three values a, b, c for the TFMF were obtained by using statistical analysis. The limits of the extreme values were determined by calculating the average + 3 standard deviations (AVG + 3σ) value. It is well known in statistics that approximately 99.7% of the data values fall within three standard deviations of the mean. All others are extreme [32]. Thus, the values that were greater than or equal than the AVG + 3σ were definitely characterized as extreme ones where the rest of them were assigned the low, medium or high linguistics based on the TFMF. For each pollutant, semi-triangular FMF was employed to assign the linguistics low and high.

After performing fuzzification of the gas emissions values for 2013, it was found that in most of the cases, the values of NO_2 were characterized as high risky, whereas in the majority of the cases for NO extremely risky values were obtained.

7 Discussion

The research effort presented in this paper manages to predict future values of air pollutants without the usage of other pollutants as inputs to the proposed model. The main idea behind the prediction model is that firstly a clustering of the preexisting dataset must occur, in order for the data to be labeled. In the next step, only the records from the "extreme" cluster (the cluster which has the biggest values for our pollutants) are taken under consideration and inserted in our prediction model. In order for the unlabeled data to gain labels (clustering creation), four unsupervised algorithms were implemented in previous efforts of our group. In this paper, the labeling process is achieved by a fuzzy semi-unsupervised algorithm: By using preexisting labels for a small amount of our dataset, the model manages to group all the records into clusters and also provide fuzzy membership values for each one of them. By checking the comparison of Tables 10 and 11, we see that the new method performs similarly with the best of the previous algorithms for each case (pollutant). In some cases, it performs even better than the previous algorithms. The main benefit though (apart from being novel) is clearly shown in the next table, where we can see the amount of records that were labeled as "extreme" by each model. We see that the fuzzy semi-unsupervised algorithm labels much more records as extreme, which makes the model more flexible in its application. Moreover, it helps us shape a better picture about air quality in Athens and allows us to predict more future values of air pollutants by considering them as extreme. Finally, the biggest advantage of the proposed algorithm as a whole is that by using semi-supervised learning, it allows us to evaluate the semi-clustering that occurred. This is a huge benefit as there is no established validation procedure for unsupervised learning. The semi-supervised algorithm is used as an inversion of the semi-unsupervised one; it performs classification on the pre-labeled dataset, and then the neural network created is used to perform classification and comparison of the classes assigned to each record with the cluster that was assigned from the semi-unsupervised sub-algorithm. This comparison is shown in Table 6. The two sub-algorithms (semi-unsupervised and semi-supervised) are not related in any manner and are implemented in different stages of the proposed algorithm. As a result, this combination of semi-unsupervised and semi-supervised learning can be used in any other similar case or individually as an evaluation method of the labeling process.

8 Conclusions and future work

This paper presents an innovative hybrid integrated approach for clustering and classification, which exploits the advantages of each sub-method. The development of the model does not require the classic initial assignment of classes for the whole dataset (which would be time, effort and resources consuming). Also the classification is done with the use of a rational, fast and timely method that requires the pre-classification only for the 10% of the data records in order to classify the rest.

Furthermore, it incorporates fuzzy logic, which adds flexibility both in terms of the functionality of the method and in terms of making the outcome easy to understand for the common people (users). Moreover, the number of features considered by the model is the minimum and there is no need for real-time feedback from sensors or other hardware devices. This not only reduces the cost, but it also reduces the runtime required.

The purpose of assigning RL to the values of the projected gas emissions is the flexible and broader understanding of the degree of air pollution risk and the effective and reliable warning of the society as a measure of public health protection policy. This is achieved by using the minimum resources in terms of cost and in terms of processing time. Everyone can be informed at no cost by using a cheap mobile device.

The proposed approach uses semi-supervised learning, which is one of the most rational and realistic machine learning methods.

The FuSSFFra algorithm allows the division of the available dataset in homogeneous sectors, based on the classification of first time seen samples, according to their distribution. This results in the enhancement of the learning process. The classification becomes faster and easier as it is performed with the minimum potential volume of resources. The FuSSFFra was successfully tested as a standalone system and in comparison with other methodologies.

Future work will include its application and testing with other datasets, not necessarily meteorological or air quality related. Also, the semi-supervised algorithm can be potentially used in a comparative mode with other classification algorithms. Finally, since the FuSSFFra works well for the city of Athens, it would be essential to run it for other areas with different climate and it would be great to try to project in the future by considering climate change for the areas under study.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Education Research Centre of Greece. http://www.kee.gr/perivallontiki/teacher6_4.html. Accessed 1 Feb 2017
2. Bougoudis I, Iliadis L, Papaleonidas A (2014) Fuzzy inference ANN ensembles for air pollutants modeling in a major urban area: the case of Athens. Eng Appl Neural Netw Commun Comput Inf Sci 459:1–14. doi:10.1007/978-3-319-11071-4_1
3. Iliadis L, Bougoudis L, Spartalis S (2014) Comparison of self organizing maps clustering with supervised classification for air

pollution data sets. Proc AIAI 436:424–435. doi:10.1007/978-3-662-44654-6_42

4. Bougoudis I, Demertzis K, Iliadis L (2016) Fast and low cost prediction of extreme air pollution values with hybrid unsupervised learning. Integr Comput Aided Eng 23(2):115–127. doi:10.3233/ICA-150505
5. Bougoudis I, Demertzis K, Iliadis L, Anezakis VD, Papaleonidas A (2016) Semi-supervised hybrid modeling of atmospheric pollution in urban centers. Commun Comput Inf Sci 629:51–63
6. Bougoudis I, Demertzis K, Iliadis L (2016) HISYCOL a hybrid computational intelligence system for combined machine learning: the case of air pollution modeling in Athens. EANN Neural Comput Appl 27:1191–1206. doi:10.1007/s00521-015-1927-7
7. Krithara A, Amini MR, Renders JM, Goutte C (2008) Semi-supervised document classification with a mislabeling error model. In: 30th European conference on IR research, ECIR 2008, advances in information retrieval, lecture notes in computer science, 4956:370–381. doi:10.1007/978-3-540-78646-7_34
8. Ashfaq RAR, Wang XZ, Huang JZ, Abbas H, He YL (2017) Fuzziness based semi-supervised learning approach for intrusion detection system. Inf Sci 378:484–497. doi:10.1016/j.ins.2016.04.019
9. Yan Y, Chen L (2011) Label-based semi-supervised fuzzy co-clustering for document categoraization. In: 8th international conference on information, communications and signal processing, (ICICS) pp 1–5. doi:10.1109/ICICS.2011.6173605
10. Zheng A, Luo L (2012) A semi-supervised fuzzy SVM clustering framework. Recent advances in computer science and information engineering, lecture notes in electrical engineering, 1:525–530. doi:10.1007/978-3-642-25781-0_78
11. Le T, Tran D, Tran T, Nguyen K, Ma W (2013) Fuzzy entropy semi-supervised support vector data description. In: Proceedings of the international joint conference on neural networks, pp 1–5. doi:10.1109/IJCNN.2013.6707033
12. Yan Y, Cui J, Pan Z (2013) Semi-supervised fuzzy relational classifier. Comput Intell Des ISCID. doi:10.1109/ISCID.2013.207
13. Benbrahim H (2011) Fuzzy Semi-supervised support vector machines. Mach Learn Data Min Pattern Recognit LNCS 6871:127–139
14. El-Zahhar MM, El-Gayar NF (2010) A semi-supervised learning approach for soft labeled data. In: Proceedings of the 10th international conference on intelligent systems design and applications (ISDA) pp 1136–1141. doi:10.1109/ISDA.2010.5687034
15. Jamalabadi H, Nasrollahi H, Alizadeh S, Araabi BN, Ahamadabadi MN (2016) Competitive interaction reasoning: a bio-inspired reasoning method for fuzzy rule based classification systems. Inf Sci 352–353:35–47. doi:10.1016/j.ins.2016.02.052
16. Cordeiro FR, Santos WP, Silva-Filho AG (2016) A semi-supervised fuzzy GrowCut algorithm to segment and classify regions of interest of mammographic images. Expert Syst Appl 65:116–126
17. Yan J, Qi W, Yue S, Zhang D, Guo D, Ma H (2016) Application of semi-supervised fuzzy kernel clustering algorithm in recognizing transformer winding's pressed state. In: ICSPCC 2016—IEEE international conference on signal processing, communications and computing, conference proceedings, 7753697, Hong Kong, China, pp 1–6. doi:10.1109/ICSPCC.2016.7753697
18. Tanaka D, Honda K, Ubukata S, Notsu A (2016) A semi-supervised framework for MMMs-induced fuzzy co-clustering with virtual samples. Adv Fuzzy Syst 2016:1–8. doi:10.1155/2016/5206048
19. Honda K, Ubukata S, Notsu A, Takahashi N, Ishikawa Y (2015) A semi-supervised fuzzy co-clustering framework and application to twitter data analysis. In: 4th international conference on informatics, electronics and vision, Fukuoka. pp 1–4. doi:10.1109/ICIEV.2015.7334057
20. Jensen R, Vluymans S, Parthaláin NM, Cornelis C, Saeys Y (2015) Semi-supervised fuzzy-rough feature selection. Lecture notes in computer science including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics 9437:185–195
21. Le T, Nguyen V, Pham T, Dinh M, Le TH (2015) Fuzzy semi-supervised large margin one-class support vector machine. Adv Intell Syst Comput 341:65–78
22. Diaz-Valenzuela I, Vila MA, Martin-Bautista MJ (2016) On the use of fuzzy constraints in semisupervised clustering. IEEE Trans Fuzzy Syst 24(4):992–999
23. Bchir O, Frigui H, Ismail MMB (2013) Semi-supervised fuzzy clustering with learnable cluster dependent kernels. Int J Artif Intell Tools 22(3):1–26. doi:10.1142/S0218213013500139
24. Sen P, Namata G, Bilgic M, Getoor L, Galligher B, Eliassi-Rad T (2008) Collective classification in network data. Adv Artif Int 29(3):93–106
25. Kecman V (2001) Learning and soft computing. MIR Press, Moscow. ISBN 9780262112550
26. Iliadis L (2007) Intelligent information systems and application in risk estimation. Stamoulis Publishing, Thessaloniki
27. Iliadis L, Papaleonidas A (2016) Computational intelligence an intelligent agents. Tziolas publications, Thessaloniki
28. Cox E (2005) Fuzzy modeling and genetic algorithms for data mining and exploration. Elsevier Science, USA
29. Anezakis VD, Dermetzis K, Iliadis L, Spartalis S (2016) Fuzzy cognitive maps for long-term prognosis of the evolution of atmospheric pollution, based on climate change scenarios: The case of Athens. Lecture notes in computer science (lecture notes in artificial intelligence and lecture notes in bioinformatics) 9875:175–186. doi:10.1007/978-3-319-45243-2_16
30. Ghosh P, Kundu K (2013) Photo-fuzzy concepts generation technique using fuzzy graph. In: Chakraborty MK, Skowron A, Maiti M, Kar S (eds) Facets of uncertainties and applications, ICFUA. Springer, Kolkata, pp 63–72
31. Cordon O, Herrera F, Hoffmann F, Magdalena L (2001) Genetic fuzzy systems evolutionary tuning and learning of fuzzy knowledge bases. Advances in fuzzy systems-applications and theory, vol 19. World Scientific Publishing, Hong Kong
32. Pukelsheim F (1994) The three sigma rule. Am Stat 48:88–91