A Recurrent Neural Network Approach to Improve the Air Quality Index Prediction

Fabio Cassano(✉), Antonio Casale, Paola Regina, Luana Spadafina, and Petar Sekulic

Omnitech Research Group, Bari, Italy
{fabio.cassano,antonio.casale,paola.regina,
luana.spadafina,petar.sekulic}@omnired.eu

Abstract. Every year, cities all over the world face the problem of the air pollution. In particular seasons, such as the winter, the levels of bad particles coming from the industrial and domestic heating systems increase the risk of pulmonary diseases. Thus, for both the city majors and citizens, it is important to understand and predict the air pollution levels in advance to safe guard the health. Modern forecasting systems are able to alert the population in advance only about the general weather condition, while the air quality information are almost not considered at all. The reasons are manifold and they mostly depend by the difficult that the modern systems have to generalize the problem and correct elaborate the data coming from the sensors. In this paper we address the problem of forecasting the bands of the different air pollutants according to the Air Quality Index in the Apulia region. Using two different Recurrent Neural Network models, we have performed two tests to prove that it is possible to predict the level of the pollutants in a specific area by using the data coming from the surrounding area. By using this approach on both the weather and air stations on the territory it is possible to have alerts many days ahead on the pollution levels.

Keywords: Air Quality Index · Recurrent Neural Network · Blind prediction

1 Introduction

The air quality level monitoring is receiving an increasing attention by researchers all over the world. Breathing clean air is essential for the human health and well-being, however, many people around the world live in places where the constant air pollution expose them to an higher risk of pulmonary disease or lung cancer [1]. Among the different pollution particles present in the air, the small one such as the $PM_{2.5}$ and the PM_{10} are considered among the most dangerous for the human health causing pulmonary cancer. In big cities close to industrial manufacture poles, the levels of those elements are so high that people moving around need to wear protective masks to safeguard their health.

P. Novais et al. (Eds.): ISAmI 2019, AISC 1006, pp. 36–44, 2020.
https://doi.org/10.1007/978-3-030-24097-4_5

There are many alert systems that allow the population to know in advance the levels of small pollutants particles, however, those are capable to predict with few advance when those level are going to exceed the limit. The overall presence of harmful particles in the air depends not only by the industrial waste, the vehicular traffic, or the urban heating systems but also by the weather conditions (such as the wind, the humidity of the air, the rain etc.). As a matter of fact, while the polluted air moves according to the direction of the wind from a city to the nearby ones expanding the discomfort and the population health-related issues, the rain "cleans" the air lowering the bad particles level. Having a good prediction of the quality of the air has manifold benefits, such as:

1. Allowing the population to be warned in time of the increasing number of pollution particles;
2. Allowing the local governments to adopt social strategies (such as limiting the traffic to a specific type of vehicles) and,
3. Allowing companies to schedule their production to avoid the excess of pollution in the air.

In this paper we address the problem of forecasting the air quality by analyzing the weather and the air pollution level of the Apulia (Italy) region. We developed two different models that, using modern Deep Learning techniques, forecast the level and the air quality (evaluated according to the Common Air Quality Index, or CAQI) considering the current weather conditions and the historical pollution data of a given place. Recurrent Neural Networks (RNNs), which are the state of the art method in sequence labelling problems, are used to predict the air quality level. We have performed two test: the first takes into account the data from 2012 to 2017 and randomly extracts the days for training and test. The second, using the same data, randomly chooses, one air station and considers it as test, while the others are used for training. In this way we prove that the algorithm is able to predict the air levels, without using the data coming from the chosen air station.

This paper is organized as follows. Section 2 describes the related works of the techniques applied to the air quality and the forecasting of the pollution. Section 3 describes the details of the proposed solution. Section 4 shows the results obtained by our tests, while Sect. 5 draws the conclusions and the future works.

2 Related Works

Nowadays Artificial Intelligence (AI) techniques are widely used in many field, from the medicine to the weather forecasting to the art. Thanks to the recent discoveries and the increasing computational power of modern personal computers, complex dataset can be elaborated in few time reaching an incredible accuracy. This can be also improved by mixing multiple approaches such as the evolutionary one and the machine learning [2,3]. Modern Machine Learning (ML) techniques such as the Artificial Neural Network, the Genetic Programming,

Deep Neural Network, etc. allow scientists from all over the world to forecast the level of harmful particles, such as $PM_{2.5}$ or the PM_{10} [4]. Those models also provide useful information on how the pollution moves from one area to another of a country. However, the forecasting accuracy rapidly decreases according the days ahead and the data used. Among the different neural network models, those with the Gated Recurrent Unit (GRU) and the Long Short Term Memory (LSTM) are used to predict values using time series [5]. However, only the usage of LSTM neural networks is currently being explored to predict the pollution in many cities [6].

To reach a good prediction level, all the ML techniques need valuable data to work with. There are many ways to retrieve environmental data: from the weather stations installed in different places of a given area, to the satellite image analysis or even by crowd-sourcing [7]. Most of the array sensor used to analyse the air are commonly installed both in the city center to track the level of bad particles from the transports, and close the industrial poles. The acquisitions are commonly scheduled in a fixed time span, however, many scientists have adopted innovative ways to get real-time data by moving through the area monitoring the level of pollutants [8]. Lastly, depending on the data acquisition rate, the ML model can be tuned to improve the prediction accuracy [9].

To easily define the overall quality of air pollutants, each nation of the world adopts a classification method called "Air Quality Index" or AQI. Its aim is to classify the air quality into different levels taking into account several parameters such as the density of micro particles like the CO_2, the O_3 etc. In general it takes into account the range of the different pollutants and classifies the level of the air into different categories. To each of those categories it is associated a "health implications" for the population [10]. This method allows a fast evaluation of the overall quality of the air, however, due to the different government laws, AQI ranges are not standardized among countries (and continents) all over the world [11]. As a matter of fact, both the level of pollutants and the number of the AQI groups depend on the nation and the standard adopted. Thus, it is possible that a "Good" AQI for a given country is considered "Moderate" for another one. This creates confusion and does not allow forecasting models developed in different countries to be easily adopted worldwide. Trying to predict the AQI index according to the different parameters is still an open question that drives researchers to use techniques such as the fuzzy logic to improve the forecast [12]. Having good results also improve people's lifestyle allowing them to spend less time at home [13].

3 The Proposed Solution

In this paper we present two tests performed on different RNN models. Both aim to the prediction of the CAQI level using data coming from weather and air stations. By using those models on each station, it is possible to predict the quality of the air and warn the population in advance for possible air pollution-related problems. In this way, each station become "smart" thanks to the integration of the those AI algorithms and the "freshness" of the acquired data.

RNN have proved to be very flexible and accurate in multiple contexts involving time series thanks to their ability to elaborate datasets. Among the different types, those that perform better use the Long Short Term Memory (LSTM) and Gated Recurrent Unit (GRU) neurons. Discussions on the differences between these two types of neurons have been explored in [14], where it has shown that the LSTM architecture performs better on some very specific ML problems (such as numeric dataset), while the GRU has a general better results on others (such as audio and some numeric one).

Table 1. Statistics of measured values. Unit, range, maximum, minimum, mean, and standard deviation values.

Element	Unit	Range	Mean	Standard deviation
NO_2	$\mu g/m^3$	$[0.0, 387.0]$	22.8	13.2
PM_{10}	$\mu g/m^3$	$[0.0, 278.0]$	35.9	23.9
O_3	$\mu g/m^3$	$[2.0, 199.0]$	92.2	25.1
Tmin	°C	$[-1.3, 41.2]$	17.9	6.9
Tmax	°C	$[-1.2, 42.3]$	18.6	7.1
Humidity	%	$[12.6, 100.0]$	66.6	17.2
Wind speed	m/s	$[0.0, 9370.0]$	2.3	37.3
Wind direction	°	$[0.0, 359.0]$	209.8	101.9
Atmospheric pressure	hPa	$[964.1, 1027.9]$	1004.8	6.9
Day of year	int	$[-1, 1]$	-	-
Day of week	int	$[1, 7]$	-	-

In order to get a good prediction accuracy, we have used two different datasets at the same time: the former related to the weather conditions, the latter related to the quality of the air in the different places of the Apulia region. Both of them provide near-real-time data, freely available on the ARPA Apulia web site[1,2]. The first dataset contains information about the geographical position of the weather stations and what kind of sensors are available. In addition, each entry contains the timestamp of the acquisition and the values coming from the sensors (such as the wind direction, the humidity etc.). The second dataset is about the general condition of the air quality. Each entry contains information about the major air pollutants including the $PM_{2.5}$, PM_{10}, CO_2, NO_2 as well as the timestamp of the acquisition. Statistical information about the dataset values and how those have been normalized are reported in Table 1. The first step to get the RNN correctly trained has been to clean the datasets by reducing the number of features the neural network had as input. As a matter of fact, the air quality

[1] http://dati.arpa.puglia.it/dataset/meteo-nrt: weather dataset, last visited 23/01/2019.

[2] http://dati.arpa.puglia.it/dataset/aria-nrt: air dataset, last visited 23/01/2019.

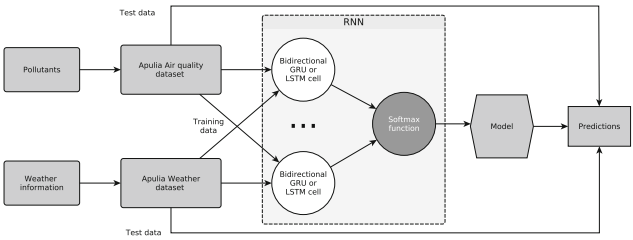


Fig. 1. The RNN model training and test schema.

station reports many pollutants value: PM_{10}, $PM_{2.5}$, NO_2, O_3, CO. However, among the available pollutants information, the Europe considers mandatory for the calculation of the CAQI only the NO_2, PM_{10}, O_3, while the $PM_{2.5}$, CO and SO_2 are optional [15]. To achieve a better accuracy in the predictions, in the proposed work, only the mandatory air pollutants are used. The workflow adopted to train the two RNN is shown in Fig. 1. The first step consists in the normalization and synchronization of all the data coming from the two datasets. As a matter of fact, both of them are near-real-time with a new acquisition each 10 min, however, coming from different type of stations they are not always perfectly synced. We conducted two different tests to prove that it is possible to reach an high accuracy in the prediction of the CAQI level, and that the RNN can be trained to forecast the CAQI level using weather and air information from different stations. The first considers the data ordered by time splitting it into two parts: the first one is used to train the RNN, while the second is used to test the network once the RNN model has been trained. The training process takes uses all the available days (from 2012 to 2017), then considers as test 5 random sequential days. In the second test it is asked to the RNN to make a "blind prediction". Firstly it has been randomly selected an air station and then it has been trained the RNN using the data coming from the others in a specific range (which by default has been set to 20 km). By calculating the angle between the selected station and the others, and analysing the weather conditions (including the wind intensity and direction), the RNN has been trained to predict the CAQI level. With the knowledge of the weather conditions of the previous days, the typical behaviour of the wind direction, the time of the year and the overall value of the pollutants, the RNN tries to estimate the value of each pollutant and the band of the CAQI. Both the RNNs have been modeled with three layers: one representing the input from the dataset, one inner layer (hidden) with 8 neuron cells (LSTM or GRU) and one as output as shown in Fig. 1.

4 Results

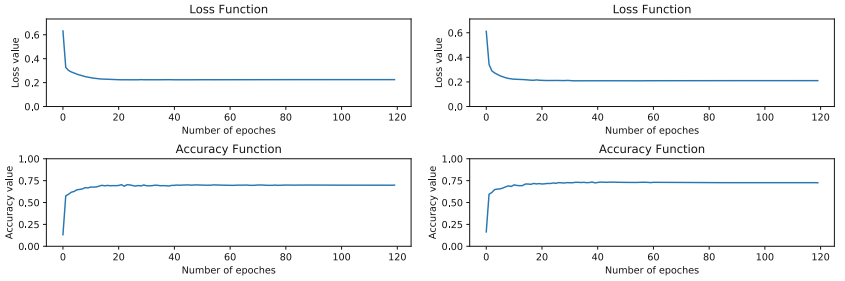
The two described dataset have been merged and fed to the RNN. In the first test performed a total of 1000 entries (almost the 80% of the dataset) have been used as training set, while 252 (almost the 20%) have been used as test

Table 2. The results (band error) of different models for daily average forecasting values of Air Quality: when the training and test samples are randomly selected (a); when the model has never seen data of predicting station (b).

Model	Forecasting	Bands	N. of days (accuracy) (a)	N. of days (accuracy) (b)
LSTM model	+1 day	±1	248(98.8%)	176(78.6%)
		±2	2(0.8%)	41(18.3%)
		±3	1(0.4%)	7(3.1%)
	+2 days	±1	246(98%)	177(79%)
		±2	4(1.6%)	44(19.7%)
		±3	1(0.4%)	3(1.3%)
	+3 days	±1	251(100%)	169(75.5%)
		±2	0(0%)	51(22.8%)
		±3	0(0%)	4(1.7%)
	+4 days	±1	247(98.4%)	175(78.1%)
		±2	4(1.6%)	45(20%)
		±3	0(0%)	4(1.8%)
	+5 days	±1	250(99.6%)	165(73.7%)
		±2	1(0.4%)	52(23.2%)
		±3	0(0%)	7(3.1%)
GRU model	+1 day	±1	247(98.4%)	170(75.9%)
		±2	4(1.6%)	44(19.6%)
		±3	0(0%)	9(4%)
	+2 days	±1	244(97.2%)	163(72.8%)
		±2	7(2.8%)	58(25.9%)
		±3	0(0%)	3(1.3%)
	+3 days	±1	249(99.2%)	159(71%)
		±2	2(0.8%)	57(25.5%)
		±3	0(0%)	7(3.1%)
	+4 days	±1	249(99.2%)	158(70.5%)
		±2	2(0.8%)	55(24.6%)
		±3	0(0%)	11(4.9%)
	+5 days	±1	251(100%)	157(70%)
		±2	0(0%)	58(25.9%)
		±3	0(0%)	8(3.6%)

set. To avoid the overfitting and the random weight initialization problem, we have repeated the training 20 times randomly choosing the training and testing data. The "blind prediction" test has a different configuration. It uses 1028

(82% of the dataset) entries as training and 224 (18% of the dataset) as test samples. The average of the final results of all trials are reported in Table 2. In the first column (Model) it is reported the model of the RNN, while the second one (Forecasting) represents the number of days in advance that the RNN have to predict. The third column (Bands) shows the CAQI value that the RNN is predicting (used like "bins" by the neural network). The last columns show how many days have been correctly predicted by the RNN when the test is performed on random data taken from the dataset or considering the blind test (respectively column "a" and column "b"). Results are excellent when it is asked the RNN to predict the CAQI level using the test data coming from the dataset. As a matter of fact, both the models are able to correctly predict with an accuracy close to the 100%. When the test data are hidden to the RNN, the accuracy in the predictions of both the models decrease. Results show that the LSTM model performs better than the GRU model trying to forecast the CAQI the band 1 one day ahead, while the opposite happens for the other bands.



(a) The loss function and the accuracy function for the LSTM network

(b) The loss function and the accuracy function for the GRU network

Fig. 2. Both the LSTM and the GRU RNN have almost the same behaviour during the training process

In Fig. 2 are shown the loss functions for both the models during the training phase of the blind prediction per epoch. Despite the number of epochs used to train the RNN, and the accuracy reached on the test data, both the loss function and the accuracy for both the LSTM and the GRU exhibit the same behaviour.

5 Conclusions and Future Works

In this paper we have faced the problem of predicting the air pollution using the RNN. Using two different datasets about the weather information and the air quality of the Apulia region, we have trained two different RNN models using the LSTM an the GRU neuron cells respectively. We have adopted this kind of network to exploit both their ability to back-propagate the error, and their classification performances on dataset involving a time series. We have used only

the CAQI mandatory pollutants and trained the RNN to predict the CAQI band for five days ahead. Two different types of experiments have been conducted: the former using as test some random data from the datasets, the latter letting the RNN to blindly predict the behaviour of a specific area knowing the behaviour of the neighbour one. Results shows that the RNN is able to predict with a very high accuracy the CAQI level of random days while the "blind prediction" results are promising.

To improve the current solution a more accurate dataset is needed. As a matter of fact, many entries from the dataset used in both the experiment were missing, thus have not been considered. To prove and tune the proposed "blind prediction" algorithm to install the model on the different air stations is needed.

Acknowledgment. The present study was developed and granted in the framework of the project: "SeVaRA" (European Community, Minister of the Economic Development, Apulia Region, BURP n. 1883 of the 24/10/2018, Id:2NQR592).

References

1. Anderson, J.O., Thundiyil, J.G., Stolbach, A.: Clearing the air: a review of the effects of particulate matter air pollution on human health. J. Med. Toxicol. **8**(2), 166–175 (2012)
2. Bevilacqua, V., Cassano, F., Mininno, E., Iacca, G.: Optimizing feed-forward neural network topology by multi-objective evolutionary algorithms: a comparative study on biomedical datasets. In: Italian Workshop on Artificial Life and Evolutionary Computation, pp. 53–64. Springer, Cham (2015)
3. Niska, H., Hiltunen, T., Karppinen, A., Ruuskanen, J., Kolehmainen, M.: Evolving the neural network model for forecasting air pollution time series. Eng. Appl. Artif. Intell. **17**(2), 159–167 (2004)
4. Feng, X., 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)
5. Che, Z., Purushotham, S., Cho, K., Sontag, D., Liu, Y.: Recurrent neural networks for multivariate time series with missing values. Sci. Rep. **8**(1), 6085 (2018)
6. Huang, C.-J., Kuo, P.-H.: A deep CNN-LSTM model for particulate matter ($PM_{2.5}$) forecasting in smart cities. Sensors **18**(7), 2220 (2018)
7. Stevens, M., D'Hondt, E.: Crowdsourcing of pollution data using smartphones. In: Workshop on Ubiquitous Crowdsourcing (2010)
8. Adams, M.D., Kanaroglou, P.S.: Mapping real-time air pollution health risk for environmental management: combining mobile and stationary air pollution monitoring with neural network models. J. Environ. Manage. **168**, 133–141 (2016)
9. Sivacoumar, R., Bhanarkar, A., Goyal, S., Gadkari, S., Aggarwal, A.: Air pollution modeling for an industrial complex and model performance evaluation. Environ. Pollut. **111**(3), 471–477 (2001)
10. To, T., Shen, S., Atenafu, E.G., Guan, J., McLimont, S., Stocks, B., Licskai, C.: The air quality health index and asthma morbidity: a population-based study. Environ. Health Perspect. **121**(1), 46–52 (2012)
11. Cheng, W.-L., Chen, Y.-S., Zhang, J., Lyons, T., Pai, J.-L., Chang, S.-H.: Comparison of the revised air quality index with the PSI and AQI indices. Sci. Total Environ. **382**(2–3), 191–198 (2007)

12. Sowlat, M.H., Gharibi, H., Yunesian, M., Mahmoudi, M.T., Lotfi, S.: A novel, fuzzy-based air quality index (FAQI) for air quality assessment. Atmos. Environ. **45**(12), 2050–2059 (2011)
13. Caivano, D., Cassano, F., Fogli, D., Lanzilotti, R., Piccinno, A.: We@ home: a gamified application for collaboratively managing a smart home. In: International Symposium on Ambient Intelligence, pp. 79–86, Springer, Cham (2017)
14. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555 (2014)
15. World Health Organization: Air quality guidelines for Europe (2000)