Prediction of Critical Air Quality Events Using Support Vector Machines and Particle Swarm Optimization

J.C. Zapata-Hernandez\textsuperscript{1}, Y.K Rojas-Idarraga\textsuperscript{1}, D.A Orrego\textsuperscript{1} and J. Murillo-Escobar\textsuperscript{1}

\textsuperscript{1} Instituto Tecnologico Metropolitano, Grupo de Investigación e Innovación Biomédica Gi2B, Medellín, Colombia

Abstract— In recent years several investigation prove the effects of air pollutants over human health, in this regard is necessary develop systems that allow people reduce the exposure to unhealthy air quality conditions. In this paper we propose a methodology to predict Critical Air Quality Events (CAQE) in Aburrá Valley based on Support Vector Machines (SVM) optimized with Particle Swarm Optimization (PSO) and a characterization scheme to assess the current and past tendencies of pollutants and weather behavior, analyzing the statistical behavior at different time intervals. We use a three stage methodology consisting in prepossessing, characterizations and CAQE prediction. The proposed method shows the better result for ozone CAQE prediction with an error of 30\%. Due to low sensitivity among the pollutants it is necessary use another machine learning technique that warranty a robust behavior working with unbalanced data.

Keywords— SVM, PSO, Air Quality, Feature Selection.

I. INTRODUCTION

The origin of urban air pollution is mainly caused due to sources of emissions lead of the human activities, such as cars, industries and domestic fuel burning [1]. Several studies have provided evidence of adverse effects associated with air pollution over human health and environment [2]. The International Agency for Research on Cancer which found sufficient evidence to classify air pollution as a leading environmental cause of cancer deaths, also according to the US agency for environmental protection, exist vulnerable groups to high levels of contamination such people with respiratory diseases as asthma, chronic obstructive pulmonary disease, bronchitis [3, 4].

The air quality information gives cities a support tool in decision-making to protect citizens, allowing a reduction in exposure to unhealthy levels of pollution, this is the case of the Po Valley (Northern Italy), they measure thirteen air pollutants concentration hourly for 13 years, they investigated the different cause as, human habits, photochemical processes of each pollutant, they processed the data statistically hourly for detect the long-term trends in relation to the changes in emission scenarios, providing direct information on the levels and trends of key pollutants [5]. Air quality models are increasingly being used not only for research but also in a working context by national weather centers and environment institutes for air quality prediction for designing emission control policies and environmental impact assessment, the regional model ensemble established in the European Union within the Mainstreaming Adaptation to Climate Change project which provides working daily air quality forecasts for Europe [6].

This paper intends to predict 24 hours early Critical Air Quality Events (CAQE) in the Aburrá Valley, using a characterization strategy that allows asses the current and past tendencies of pollutants and weather behavior through the analysis of the statistical behavior at different time intervals. The predictions was perform using a hybrid machine learning strategy based on Support Vector machines and Particle Swarm Optimization.

II. MATERIALS AND METHODS

A. Support Vector Machines

Given a training vectors $x_i \in \mathbb{R}^n, i = 1, \ldots, l$, in two classes and an indicator vector $y \in \mathbb{R}^l$ such that $y_i \in \{1,-1\}$. C-SVC [7] solves the following dual optimization problem which is the key to extend SVM for nonlinear functions:

$$\max \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j \phi(x_i, x_j)$$

Subject to: $$\begin{cases} 0 \leq \alpha_i \leq C \\ \sum_{i=1}^{l} \alpha_i y_i = 0 \end{cases}$$

where $\phi(x_i)$ is a kernel function that maps $x_i$ into a higher-dimensional space and $C > 0$ is the cost parameter. In this sense the optimization problem is expressed in terms of the lagrange multipliers $\alpha_i$ only, finally the decision function is $\text{sgn}(\sum_{i=1}^{l} y_i \alpha_i \phi(x_i, x_j) + b)$. 

© Springer Nature Singapore Pte Ltd. 2017
I. Torres et al. (eds.), VII Latin American Congress on Biomedical Engineering CLAIB 2016, Bucaramanga, Santander, Colombia, October 26th -28th, 2016, IFMBE Proceedings 60, DOI: 10.1007/978-981-10-4086-3_153
B. Particle Swarm Optimization

Swarm-based algorithms emerged as a powerful family of optimization techniques, inspired by the collective behavior of social animals. In particle swarm optimization (PSO) the set of candidate solutions to the optimization problem is defined as a swarm of particles flowing through the parameter space defining trajectories which are driven by their own and neighbors’ best performances [8], the algorithm 1 resume the employed PSO.

**Inputs:** \( f(x) \), stop criterion, \( \alpha, \beta, \theta \) & \( n \)

**Output:** \( g^* \)

**Initialization**: positions \( x_i \) and speed \( v_i \) of the \( n \) particles

\[
x_i^t = x_i, \quad v_i^t = f(x_i)
\]

\[
g^* = \arg\min f(x_i), \quad v_i^t = f(g^*)
\]

while stop criterion do
  for every \( i \) do
    \[
    v_i^{t+1} = \theta v_i^t + \alpha \varepsilon_1 \cdot [g^* - x_i^t] + \beta \varepsilon_2 \cdot [x_i^t - x_i^t]
    \]
    if \( f(x_i^{t+1}) < v_i \) then
      \[
      x_i^{t+1} = f(x_i^{t+1}), \quad x_i^t = x_i^{t+1}
      \]
    end
    if \( f(x_i^{t+1}) < v_i \) then
      \[
      g^* = f(x_i^{t+1}), \quad g^* = x_i^{t+1}
      \]
  end
end

\( \alpha \) & \( \beta \) constants of group and individual behavior \( f(x) \) objective function, \( \theta \) inertia parameter, \( \varepsilon_1 \) & \( \varepsilon_2 \) Gaussian random variables, \( v_i \) particle speed, \( x_i \) particle position, \( x_i^t \) best particle position, \( g^* \) best position reached at time \( t \)

**Algorithm 1: PSO**

C. Proposed Methodology

The fig 1 shows the proposed methodology to predict CAQE 24 hours early, initially a pre-processing stage was carried out for outliers detection, database selection and CAQE identification, moreover we conduct the data characterization and finally we use SVM optimized with PSO (SVM-PSO) to predict critical air quality events in urban and suburban areas.

D. Data Base Description

For this study we used the information provided by the Air Quality Laboratory (CALAIRE) of the National university of Colombia (UNAL). The data correspond to five stations in different sites of The Aburra Valley including Suburban and urban areas, in the urban areas we find 3 stations: Museum of Antioquia (MED-MANT), UNAL Medellín Campus (MED-NAL) and Itagui’s House of Justice (ITA - HJ), the suburban areas include San Buenaventura university (BEL-USBV) and Liceo Council Itagui (ITA-CONC).

They collect this data between January 2013 and December 31 of the same year, including measures hour after hour of six weather variables (air speed (WS), air direction (WD), air temperature (AT), humidity relative (AH), solar radiation (SR)) and six pollutants (nitrogen monoxide (NO), nitrogen dioxide (NO2), Ozone (O3), Carbon monoxide (CO), particulate matter 2.5 \( \mu m \) (PM2.5), particulate matter 10 \( \mu m \) (PM10)). Al-though to clarify that monitoring of these six pollutants were not taken simultaneously in the five stations, but was taking in combinations of three or four of these. Allowing in this way to get a representation space between ten and twelve features and a total samples of 8760.

E. Data Preprocessing

The used data present missing values and outliers, occasioned for damage or maintenance in the sensors, taking into account that this values can affect how mathematical or statistics methods perform, we conduct a pre-processing stage. First we make the outliers detection using a box plot and then replacing as a missing values. Finally, the samples with missing data was removing. In order to make a comparative analysis between a urban station that is characterized by high levels of pollutants and a suburban station where its levels of pollutants are moderate, we just use the data base with the lower number of missing values. Then, we evaluated the pollutants to identify if their concentration levels corresponding to a CAQE, we get the mean (\( \mu \)) and the standard deviation (\( \sigma \)) of every pollutant, if the concentration of the pollutant is greater than \( \mu + 0.8\sigma \) was labeled as CAQE.

F. Characterization

With the aim to shape a feature space of effective representation for CAQE prediction 24 hours early, we characterize the pollutants concentration and the weather variables in the time domain, in this way we calculate features that allows us to assess the current and past tendencies of pollutants and weather, the first group of features consisting in the measure of the pollutants concentration and weather variables 24 hours before of the prediction, for evaluate the past tendencies, we split the analysis in pollutants and weather variables, given that the pollutants concentration varies mainly as function of the human activity, we get the pollutant concentration the same hour and day in the past seven weeks and calculate the mean, maximum
and minimum concentration, in other hand the weather varies as function of region’s climatic patterns, for this reason we get the mean, maximum and minimum of the weather variable in several time intervals before the CAQE prediction (12 Hours, 24 Hours and 7 days). Later we make feature selection using an algorithm based on Fuzzy-Rough set theory propose in [9], with the aim of reduce the problems associated with high dimensional space.

G. Prediction

For prediction, we performed an algorithm of classification based on SVM-PSO. In this step, we randomly split the data in three groups: training (60 %), validation (20%) and test (20%). The SVM learning process is conducted using the training data and the parameters C and γ is given for the PSO algorithm, is important note that PSO choose the setting that maximize the accuracy over the validation data, in this process we executed 200 iterations using 15 particles, θ = 0.9 and α, β = 2. Finally we assessed the obtained model through the test data, measuring the accuracy, sensitivity and specificity in the classification process.

III. RESULTS AND DISCUSSION

A. Data Preprocessing

In this stage we realize the outlier detection, missing value removing, data base selection and CAQE identification. The urban station MED-MANT, MED-NAL and ITA-HJ has a proportion missing values of 12.7 %, 15.5 % and 15.8 % respectively, for suburban areas ITA-CONC and BEL-USB has 10.8 % and 12.7 % of missing values. The selected urban station was MED-MANT and the sub-urban station were ITA-CONC.

The monitor station MED-MANT measure the concentration of four pollutants: NO, NO₂, PM₂.₅ and CO, now we identify the CAQE for every pollutant and find that for PM₂.₅ was 3781 CAQE this correspond to the 43.2% of samples, the same way for NO was the 44.7%, NO₂ 40.7% and for CO the 34.4%. The ITA-CONC station measure three pollutants PM₂.₅, PM₁₀ and O₃, for PM₂.₅ we found a 39.4% of CAQE, PM₁₀ has a 39.7% and O₃ with 44.6%. The mean CAQE portion for the 7 pollutants is 42%, this evidence that balance between classes is not ideal and could lead in a poor performance due to increase the probability that the classification algorithm label the test instance in the group with more samples.

B. Characterization and Feature Selection

Following the proposed characterization scheme, we find a total of 47 features split in three groups (table 1). The feature selection stage was conducted to find the optimal subset of features for CAQE prediction in each pollutant, in this regard for predict CAQE in the urban area of MED-NAL associated with NO we employ only 11 features, for NO₂ 12 features, CO 8 features and PM₂.₅ 11 features, in the suburban area ITA-CONC for PM₂.₅, PM₁₀ and O₃ the algorithm select a total of 11 different features for every pollutant.

<table>
<thead>
<tr>
<th>First group</th>
<th>Second group</th>
<th>Third group</th>
</tr>
</thead>
<tbody>
<tr>
<td>WD ≥</td>
<td>M-PM₂.₅-24h★</td>
<td>M-AT-7w ★</td>
</tr>
<tr>
<td>AT ≥</td>
<td>M-PM₂.₅-7w</td>
<td>Max-AT-7w</td>
</tr>
<tr>
<td>AH ≥</td>
<td>Max-PM₂.₅-7w ≥</td>
<td>Min-AT-7w</td>
</tr>
<tr>
<td>SR ≥</td>
<td>Min-PM₂.₅-7w ≤</td>
<td>M-AT-24h</td>
</tr>
<tr>
<td>NO ≤</td>
<td>M-NO-24h ≤</td>
<td>Max-AT-24h</td>
</tr>
<tr>
<td>NO₂ ≥</td>
<td>M-NO-7w ≥</td>
<td>Min-AT-24h</td>
</tr>
<tr>
<td>PM₂.₅ ≥</td>
<td>Max-NO-7w ≥</td>
<td>M-SR-12h ≥</td>
</tr>
<tr>
<td>CO ≥</td>
<td>Min-NO-7w ≥</td>
<td>Max-SR-12h ≥</td>
</tr>
<tr>
<td>PM₁₀ ≥</td>
<td>M-NO₂-24h ≥</td>
<td>Min-SR-12h</td>
</tr>
<tr>
<td>O₃ ≥</td>
<td>M-NO₂-7w ≥</td>
<td>M-AT-7w ≥</td>
</tr>
<tr>
<td>WS ≥</td>
<td>Max-NO₂-7w ≥</td>
<td>Max-AT-7w ≥</td>
</tr>
<tr>
<td></td>
<td>Min-NO₂-7w ≤</td>
<td>Min-AT-7w ≤</td>
</tr>
<tr>
<td></td>
<td>M-CO-24h ≥</td>
<td>Max-CO-7w ≥</td>
</tr>
<tr>
<td></td>
<td>M-CO-7w ≥</td>
<td>Min-CO-7w ≥</td>
</tr>
<tr>
<td></td>
<td>M-PM₁₀-7w ≥</td>
<td>Max-PM₁₀-7w ≥</td>
</tr>
<tr>
<td></td>
<td>Min-PM₁₀-7w ≤</td>
<td>M-O₃-24h ≤</td>
</tr>
<tr>
<td></td>
<td>Min-O₃-7w ≤</td>
<td></td>
</tr>
</tbody>
</table>

MED-NAL. ⊗ selected for NO, ⊘ selected for NO₂, ★ selected for CO
ITA-CONC. ⊙ selected for O₃, selected ⊗ for PM₂.₅, selected ⊘ for PM₁₀

Figure 1. Proposed Methodology


C. Critical Air Quality Events Prediction

Using the proposed SVM-PSO algorithm we conduct the CAQE prediction in urban and suburban areas, the selected parameter for PSO in urban areas pollutants were $C = 88.3$ and $\gamma = 2.9$ for NO CAQE prediction, for $NO_2$ $C = 0.74$ and $\gamma = 10$, for $PM_{2.5}$ $C = 233.3$ and $\gamma = 3.6$, for CO $C = 0.8$ and $\gamma = 5.1$. In suburban areas PSO gave $C = 1000$ and $\gamma = 0.09$ for $O_3$, $C = 779.14$ and $\gamma = 0.07$ for $PM_{10}$ and finally for $PM_{2.5}$ we obtain $C = 10.25$ and $\gamma = 0.78$.

Table 2 show the system performance for urban and suburban areas, the best result in urban areas was obtained for the NO and the poorest were reach for $PM_{2.5}$, while in suburban areas the best performance where reach for $O_3$, the poorest agree with urban area for $PM_{2.5}$. The sensitivity in the CAQE detection in the 7 contaminats is low, due to the unbalance between classes, moreover the feature selection stage in $PM_{2.5}$ for MED-NAL show that the proposed features are uncorrelated with CAQE given that only one feature was selected.

<table>
<thead>
<tr>
<th>Pollutant</th>
<th>Spe</th>
<th>Sen</th>
<th>Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>NO</td>
<td>0.72</td>
<td>0.64</td>
<td>0.69</td>
</tr>
<tr>
<td>$NO_2$</td>
<td>0.86</td>
<td>0.38</td>
<td>0.66</td>
</tr>
<tr>
<td>$PM_{2.5}$</td>
<td>1</td>
<td>0</td>
<td>0.56</td>
</tr>
<tr>
<td>CO</td>
<td>0.85</td>
<td>0.44</td>
<td>0.68</td>
</tr>
<tr>
<td>$O_3$</td>
<td>0.8</td>
<td>0.59</td>
<td>0.7</td>
</tr>
<tr>
<td>$PM_{10}$</td>
<td>0.89</td>
<td>0.2</td>
<td>0.63</td>
</tr>
<tr>
<td>$PM_{2.5}$</td>
<td>0.89</td>
<td>0.2</td>
<td>0.61</td>
</tr>
</tbody>
</table>

IV. Conclusion

This paper attempts to predict Critical Air Quality Events (CAQE) using Support Vector Machines (SVM) Optimized with Particle Swarm Optimization (PSO) through a characterization scheme to asses the current and past tendencies of pollutants and weather behavior through statistical features, obtaining a low performance in terms of sensibility and accuracy. The unbalance between classes lead in low sensibility since most of the label correspond to normal air quality conditions, this is confirmed through the high specificity, for these reason is necessary employ a machine learning techniques developed to work with unbalanced data. For some pollutants the feature selection algorithm chose a few features in the case of $PM_{2.5}$ in MED-NAL only one feature this can suggest that the proposed set of features are not so related to CAQE and is necessary extend the analysis to non linear features.

Conflict of interest

The authors declare that they have no conflict of interest.

Acknowledgements

We would like to thank the Air Quality Laboratory of the National university of Colombia and Metropolitan Area of the Aburrá Valley for provide the data used in this study.

References


Author: J. Murillo-Escobar
Institute: Instituto Tecnológico Metropolitano
Street: Calle 73 No 76A-354
City: Medellín
Country: Colombia
Email: jpmurillo22@gmail.com