Analysis and Prediction of Air Quality Data with the Gamma Classifier

Cornelio Yáñez-Márquez, Itzamá López-Yáñez,
and Guadalupe de la Luz Sáenz Morales

IPN Centro de Investigación en Computación
Juan de Dios Bátiz s/n esq. Miguel Othón de Mendizábal
Unidad Profesional Adolfo López Mateos
Del. Gustavo A. Madero, México, D.F., México
cyanez@cic.ipn.mx, ilopezyb05@ipn.mx,
gsaenza08@sagitario.cic.ipn.mx

Abstract. In later years, different environmental phenomena have attracted the attention of artificial intelligence and machine learning researchers. In particular, several research groups have applied genetic algorithms and artificial neural networks to the analysis of data related to atmospheric and environmental sciences. In the current work, the results of applying the Gamma classifier to the analysis and prediction of air quality data related to the Mexico City Air Quality Metropolitan Index (IMECA in Spanish) are presented.

Keywords: Air quality forecast, gamma classifier.

1 Introduction

Environmental topics have gained the attention of large portions of population. In different languages and through diverse means, civil associations launch campaigns for people awareness of the importance of protecting the environment [1], [2], even attracting governments active participation [3]-[6].

Several techniques of artificial intelligence have been applied to the analysis of environmental data, such as artificial neural networks [7]-[9] and Support Vector Machines [10]. In this paper the Gamma classifier [11] is applied to the analysis and prediction of the Mexico City Air Quality Metropolitan Index (*Índice Metropolitano de la Calidad del Aire*, IMECA in Spanish) related data.

The rest of the paper is organized as follows: the air quality data, SIMAT and IMECA, are described in section 2, while section 3 is dedicated to the Gamma classifier. Section 4 contains the main proposal of this work, and in section 5 the experimental results are discussed. Conclusions are shown in section 6.

2 SIMAT and IMECA

The Mexico City Atmospheric Monitoring System (*Sistema de Monitoreo Atmosférico*, SIMAT in Spanish) is tightly coupled with the evolution of the

J. Ruiz-Shulcloper and W.G. Kropatsch (Eds.): CIARP 2008, LNCS 5197, pp. 651–658, 2008.

Table 1. IMECA and its implications for Health

IMECA	Condition	Effects on Health
0-50: green	Good	Suitable for conducting outdoor activities
51-100: yellow	Regular	Possible discomfort in children, the elderly and people with illnesses
101–150: orange	Bad	Adverse health effects on the population, particularly on children and older adults with cardiovascular and / or respiratory illnesses such as asthma
151–200: red	Very Bad	Greater adverse health effects on the population, particularly on children and older adults with cardiovascular and / or respiratory illnesses such as asthma
>200: purple	Extremely Bad	Adverse health effects in the general population. Serious complications may present in children and older adults with cardiovascular and / or respiratory illnesses such as asthma

Table 2. IMECA Calculation for sulfur dioxide (SO2)

IMECA Interval	Concentration Intervals (ppm)	Equations
0-50: green	0-0.065	$\mathrm{IMECA}[SO_2]= \frac{\mathrm{con}[SO_2]*100}{0.13}$
51-100: yellow	0.066-0.130	
101–150: orange	0.131-0.195	
151–200: red	0.196-0.260	
>200: purple	>0.260	

Mexican capital, and with the problems inherent to its development. The information herein presented is taken from [12].

SIMAT is committed to operate and maintain a trustworthy system for the monitoring of air quality in Mexico City, as well as analyzing and publishing this information. It is made up by four specialized subsystems (RAMA, REDMA, REDMET, and REDDA; see [12] for the acronyms meanings), one Atmospheric Monitoring Mobile Unit, and a Calibration Standards Transfer Laboratory.

The IMECA is a reference value for people to be aware of the pollution levels prevalent in any zone, in a precise and timely manner; in order to take appropriate protection measures. When the IMECA of any pollutant is greater than 100 points, its concentration is dangerous for health and, as the value of IMECA grows, the symptoms worsen, as can be seen in table 1.

Generating the IMECA is one of the primordial tasks of SIMAT. Since July 1st., 1998, the IMECA has been transmitted 24 hours every day to different electronic and printed communication media.

In November 2006, the *Gaceta Oficial del Distrito Federal* published the *Norma Ambiental para el Distrito Federal* NADF-009-AIRE-2006 [13], which states the specifications for computing the IMECA for the criteria pollutants, such as: O_3, NO_2, SO_2, CO, PM10 and PM2.5. For each of the criteria pollutants, this norm states equations for calculating the corresponding IMECA, from the concentration data in parts per million (ppm). For instance, table 2 shows how the sulfur dioxide (SO_2) IMECA is calculated.

3 The Gamma Classifier

This pattern classifier, of recent proposal, has shown some very promising results. The following discussion is strongly based on [11].

The basis of the Gamma classifier is the gamma operator, hence its name. In turn, the gamma operator is based on the alpha, beta, and u_β operators and their properties, in particular when dealing with binary patterns coded with the modified Johnson-Möbius code. Also, it is important to define the sets A and B, since they are used throughout this work. Thus, $A = \{0, 1\}$ and $B = \{0, 1, 2\}$.

3.1 Preliminaries

The alpha and beta operators are defined in tabular form as shown in table 3.

Table 3. Definition of the Alpha and Beta operators

$\alpha : A \times A \to B$	
x y	$\alpha(x, y)$
0 0	1
0 1	0
1 0	2
1 1	1

$\beta : B \times A \to A$	
x y	$\beta(x, y)$
0 0	0
0 1	0
1 0	0
1 1	1
2 0	1
2 1	1

The unary operator u_β, which receives as input an n-dimensional binary vector $\mathbf{x}$, and outputs a non-negative integer number, is calculated as follows:

$$u_\beta(\mathbf{x}) = \sum_{i=1}^{n} \beta(x_i, x_i) \tag{1}$$

The modified Johnson-Möbius code converts a set of real numbers into binary representations by: (a) substracting the minimum from each number, leaving only non-negative reals; (b) scaling up the numbers, leaving only non-negative integers (truncating the left decimals if necessary); (c) concatenating $e_m - e_j$ zeros with e_j ones, where e_m is the greatest non-negative integer number (already shifted, truncated and scaled) to be coded, and e_j is the current non-negative integer number to be coded.

The generalized gamma operator γ_g, which takes as input two binary patterns $\mathbf{x} \in A^n$ and $\mathbf{y} \in A^m$, with $n, m \in \mathbb{Z}^+$, $n \leq m$, and a non-negative integer number θ; and gives a binary number as output; can be computed as:

$$\gamma_g(\mathbf{x}, \mathbf{y}, \theta) = \begin{cases} 1 \text{ if } m - u_\beta[\alpha(\mathbf{x}, \mathbf{y}) \bmod 2] \leq \theta \\ 0 \text{ otherwise} \end{cases} \tag{2}$$

3.2 The Gamma Classifier Algorithm

Let $k, m, n, p \in \mathbb{Z}^+$; $\{\mathbf{x}^\mu \,|\mu = 1, 2, \ldots, p\}$ be the fundamental pattern set with cardinality p, where $\forall\mu\ \mathbf{x}^\mu \in \mathbb{R}^n$, and let $\mathbf{y} \in \mathbb{R}^n$ be an n-dimensional real-valued pattern to be classified. It is assumed that the fundamental set is partitioned into m different classes, each class having a cardinality k_i, $i = 1, 2, \ldots, m$, thus $\sum_{i=1}^{m} k_i = p$. In order to classify $\mathbf{y}$, these steps are followed:

1. Code the fundamental set with the modified Johnson-Möbius code, obtaining a value $e_m = \bigvee_{i=1}^{p} x_j^i$ for each component.
2. Compute the stop parameter $\rho = \bigwedge_{j=1}^{n} e_m\,(j)$.
3. Code $\mathbf{y}$ with the modified Johnson-Möbius code, using the same parameters used with the fundamental set. If any y_j is greater than the corresponding $e_m\,(j)$, the γ_g operator will use such y_j instead of m.
4. Transform the index of all fundamental patterns into two indices, one for the class they belong to, and another for their position in the class (i.e. $\mathbf{x}^\mu$ which belongs to class i becomes $\mathbf{x}^{i\omega}$).
5. Initialize θ to 0.
6. Do $\gamma_g\left(\mathbf{x}_j^{i\omega}, \mathbf{y}_j, \theta\right)$ for each component of the fundamental patterns.
7. Calculate $c_i = \frac{\sum_{\omega=1}^{k_i}\sum_{j=1}^{n}\gamma_g\left(\mathbf{x}_j^{i\omega}, \mathbf{y}_j, \theta\right)}{k_i}$ for each class.
8. If there is more than one maximum among the different c_i, increment θ by 1 and repeat steps 6 and 7 until there is a unique maximum, or the stop condition $\theta \geq \rho$ is fulfilled.
9. If there is a unique maximum, assign $\mathbf{y}$ to the class corresponding to such maximum: $C_y = C_j$ such that $\bigvee_{i=1}^{m} c_i = c_j$.
10. Otherwise, assign $\mathbf{y}$ to the class of the first maximum.

4 Proposed Application

The main proposal of the current paper is to apply the Gamma classifier to environmental data taken from SIMAT databases (in particular the RAMA database), to automatically predict future values. For this, all the samples taken at Tacuba monitoring station for SO_2 in year 2001 are used to form the fundamental set. For the testing set, the samples taken at the same station during February 2002 are used. Each pattern is made up by n successive samples, concatenated each after the other. As the class for such pattern, the $n+1$-th sample is used.

There are three values of interest to this application. First, the directly predicted value: SO_2 concentration, in ppm. Second, the IMECA value calculated from such concentration, in points. And finally, the IMECA level or interval, as a category: good, regular, bad, very bad, and extremely bad. For computational purposes, these categories were coded as a number: $1, 2, 3, 4, 5$ respectively.

5 Experimental Results

As mentioned in the previous section, both the fundamental set and the testing set were formed with data taken from the RAMA database for SO_2, containing hourly samples of concentration measured in ppm. The fundamental set contains a total of 8040 samples; while the testing set contains a total of 672 samples.

With these data, input patterns of 10 samples were formed ($n = 10$). While the value of n can be arbitrarily chosen, 10 gave the best results in preliminary tests. The output patterns (i.e. the class) were taken from the sample following the last sample in the pattern. Notice that sample values are presented in thousandths of ppm; thus, they were scaled by 1000 in order to handle integers. Also, there are as many classes as different output patterns; however, the current task being prediction, not classification, this number is of little interest.

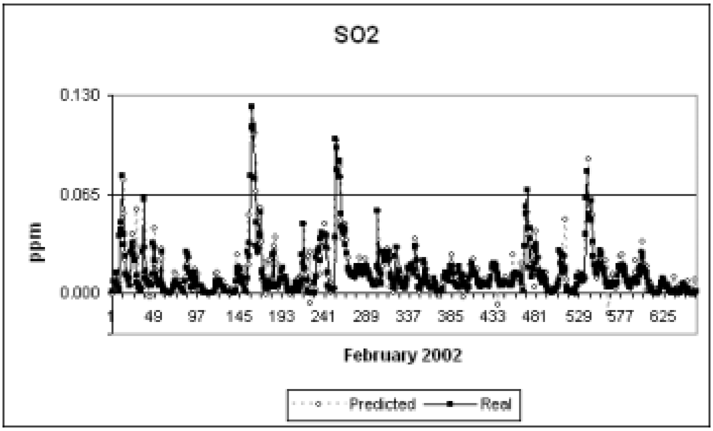

Fig. 1. Predicted values *vs* real values for SO_2 concentration

With these considerations, the fundamental set is made up of 8029 associations (x, y) and the testing set is made up of 661 associations (x, y), both with input patterns $x \in \mathbb{R}^{10}$ and output patterns $y \in \mathbb{R}$.

Once trained with the fundamental set, the Gamma classifier is presented with the testing set, obtaining the predicted values of SO_2 concentration (see figure 1). With these values and using the equations shown in table 2, the IMECA value is computed (see figure 2). Finally, the obtained IMECA values are classified inside a range in order to obtain the corresponding IMECA level. Some illustrative examples of these results are shown in table 4.

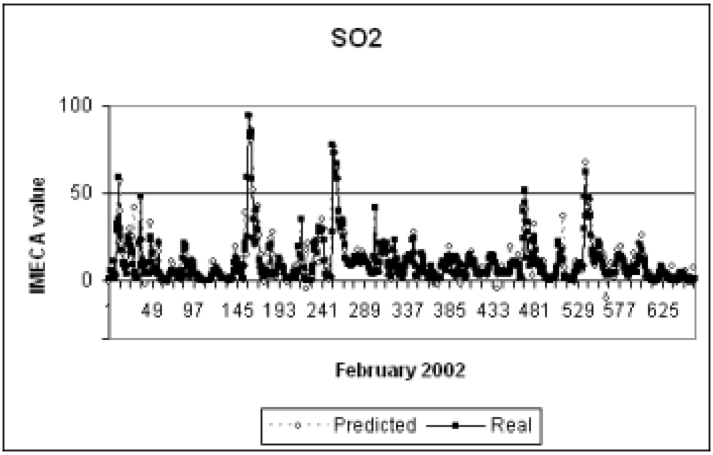

Fig. 2. Predicted values *vs* real values for SO_2 IMECA value

Table 4. Examples of results; (P) Predicted, (R) Real, (E) Error

Sample	Concentration			IMECA Value			IMECA Level		
	P	R	E	P	R	E	P	R	E
Feb. 1 14:00	0.008	0.008	0.000	6	6	0	1	1	0
Feb. 10 10:00	0.012	0.025	0.013	9	19	−10	1	1	0
Feb. 22 20:00	0.048	0.016	0.032	37	12	25	1	1	0

Notice how, although one of the largest errors exhibited in concentration and IMECA value prediction is seen on Feb. 22 20:00 (0.032 ppm and 25), the IMECA level is predicted correctly.

Two quantitative measures of the performances shown by the Gamma classifier on this application were used. On one side, the Rooted Mean Square Error (RMSE), which is a widely used measure of performance and is calculated as shown in equation 3. On the other side, the bias, which can be calculated by following 4, is used to describe how much the system is underestimating or over estimating the results. For both equations, P_i is the i-th predicted value and O_i is the i-th original (real) value.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(P_i - O_i)^2} \tag{3}$$

$$Bias = \frac{1}{n}\sum_{i=1}^{n}(P_i - O_i) \tag{4}$$

The RMSE for concentration is 0.009218, while for IMECA value it is 7.093710 and 0.116686 for IMECA level. The bias for concentration, IMECA value, and level are, respectively: 0.270, 202, and −1.

It is clear that, while there are some significant errors in concentration prediction, these are few and not that large; which in turn is reflected in the IMECA value performance, with expected error of less than 10 points. This is also reflected in the IMECA level, where the RMSE indicates an expected error of about 1/10 of level. The bias, on the other hand, shows some interesting behavior: while it has positive values for both concentration and IMECA value, the IMECA level has a negative bias. It is also noteworthy that the bias for concentration is 0.270, which is rather low considering that these results encompass 661 test patterns. It is also remarkable that for IMECA level, there are only 9 errors out of 661 instances.

In tables 5 and 6 a comparison with related works is presented. Notice that the results obtained with the Gamma classifier are quite competitive.

Table 5. Comparison of related results (SIMAT database) in RMSE, given for IMECA value / IMECA level; NA indicates a not available value

Algorithm Used	Pollutants Considered	Size of Training / Testing Sets	Performance
Bayesian network [7]	O_3 (ppm)	400 / 200	26.8 / 10
Neural network [7]	O_3 (ppm)	400 / 200	19.4 / NA
C4.5 [7]	O_3 (ppm)	400 / 200	21.4 / NA
Gamma classifier	SO_2 (ppm)	8040 / 672	7.09 / 0.12

Table 6. Comparison of related results (diverse databases) in RMSE, given for pollutant concentration; NA indicates a not available value, ppb means parts per billion

Algorithm Used	Pollutants Considered	Size of Training / Testing Sets	Performance
Neural network [8]	O_3 (μg/m^3)	613 / 105	15
Neural network [9]	O_3 (ppb)	NA / 1343	9.43
		NA /2367	13.79
Online SVM [10]	SO_2 (μg/m^3)	240 / 168	12.96, 10.90
CALINE3 [14]	PM_{10}, $PM_{2.5}$ (μg/m^3)	~120	88, 55
Gamma classifier	SO_2 (ppm)	8040 / 672	0.009218

6 Conclusions and Future Work

In this paper, the utility of applying the Gamma classifier to the prediction of unknown environmental data has been experimentally shown. More specifically, the hourly concentration of SO_2, its corresponding IMECA values, as well as the resulting IMECA levels, taken from the RAMA database, were analyzed.

The experimental results show a low error when compared to the data being predicted, even more so the IMECA level data. However, it is noteworthy that most significant errors occur when the graph of the data changes direction, implying a quite likely venue of improvement.

Acknowledgements. The authors would like to thank the Instituto Politécnico Nacional (Secretaría Académica, COFAA, SIP, and CIC), the CONACyT, and SNI for their economical support to develop this work.

References

1. Toepfer, K., et al.: Aliados Naturales: El Programa de las Naciones Unidas para el Medio Ambiente y la sociedad civil (in Spanish). UNEP-United Nations Foundation (2004)
2. Hisas, L., et al.: A Guide to the Global Environmental Facility (GEF) for NGOs. UNEP-United Nations Foundation (2005)
3. United Nations: Rio Declaration on Environment and Development (1992)
4. United Nations: Kyoto Protocol to The United Nations Framework Convention on Climate Change (1997)
5. Secretaría de Comercio y Fomento Industrial: Determiaci ón de Neblina de Ácido Fosfórico en los Gases que Fluyen por un Conducto (in Spanish). Mexican Norm NMX-AA-090-1986. Mexico (1986)
6. Web del Departamento de Medio Ambiente y Vivienda de la Generalitat de Cataluña (in Spanish) (2007), http://mediambient.gencat.net/cat
7. Sucar, L.E., Pérez-Brito, J., Ruiz-Suárez, J.C., Morales, E.: Learning Structure from Data and Its Application to Ozone Prediction. Applied Intelligence 7(4), 327–338 (1997)
8. Dutot, A., Rynkiewicz, J., Steiner, F.E., Rude, J.: A 24-h forecast of ozone peaks and exceedance levels using neural classifiers and weather predictions. Environmental Modelling and Software 22(9), 1261–1269 (2007)
9. Salazar-Ruiz, E., et al.: Development and comparative analysis of tropospheric ozone prediction models using linear and artificial intelligence-based models in Mexicali, Baja California (Mexico) and Calexico, California (US). Environmental Modelling and Software 23(8), 1056–1069 (2008)
10. Wang, W., Men, C., Lu, W.: Online prediction model based on support vector machine. Neurocomputing 71(4-6), 550–558 (2008)
11. Yáñez, L., Itzamá: Clasificador Automá tico de Alto Desempeño (in Spanish). M.Sc. Thesis. National Polytechnics Institute, Computers Research Center, Mexico (2007)
12. Sistema de Monitoreo Atmosférico de la Ciudad de Mé xico: IMECA (in Spanish) (2007), http://www.sma.df.gob.mx/simat/pnimeca.htm
13. Norma Ambiental para el Distrito Federal (in Spanish). Gaceta Oficial del Distrito Federal, XVI Epoch (2006)
14. Gokhale, S., Raokhande, N.: Performance evaluation of air quality models for predicting PM10 and PM2.5 concentrations at urban traffic intersection during winter period. Science of the Total Environment 394(1), 9–24 (2008)