Predicción de la calidad del aire de la Ciudad de México basado en minería de datos

Nahun Loya[1] , Hortensia Reyes[2], Yuridiana Alemán[1] y Helena Gómez-Adorno[1]
[1] Facultad de Ciencias de la Computación, [2]Facultad de Ciencias Físico-Matemáticas,
[1,2]Benemérita Universidad Autónoma de Puebla.
72180 Puebla, México.
nahun.loya@gmail.com, hreyes@fcfm.buap.mx, yuridiana.aleman@gmail.com,
helena.adorno@gmail.com
(Paper received on August 10, 2012, accepted on August 24, 2012)

Resumen. En el presente artículo se hace una revisión de los principales algoritmos de clasificación supervisada, con el objetivo de encontrar modelos para pronosticar los niveles de calidad del aire en base a un conjunto de atributos (4 químicos y 4 atmosféricos), se usan los datos del Sistema de Monitoreo Atmosférico de la Ciudad de México (SIMAT) correspondientes a las mediciones horarias de los años 2010 a 2011 considerando las estaciones meteorológicas: Pedregal, Tlalnepantla y Xalostoc. Los modelos obtenidos son usados para predecir la calidad del aire, éstos tienen una precisión de hasta 92.14%.

Palabras clave: Árboles de decisión, C4.5, Pronóstico de la calidad del aire, SIMAT, Redes neuronales, Decision Stump, Random Trees, Random Forest, ozono.

1 Introducción

Las grandes ciudades como Los Ángeles, Tokio, Moscú y la Ciudad de México presentan problemas de contaminación ambiental, éstas monitorean la calidad del aire en la tropósfera con el objetivo de documentar el problema de ozono y medir el progreso de la reducción de concentraciones emitidas por los habitantes [1].

El ozono es una molécula que está compuesta por tres átomos de oxígeno. En la tropósfera terrestre es un poderoso oxidante que reacciona rápidamente con otros compuestos químicos. El ozono penetra a través de las membranas celulares causando efectos nocivos en la salud. Cuando una persona se expone a altas concentraciones sufre trastornos en el tracto respiratorio. Estudios epidemiológicos han encontrado asociaciones entre los niveles diarios de ozono y la mortalidad.

Es importante encontrar formas de documentar, analizar, modelar y pronosticar el fenómeno de la contaminación.

El objetivo de este trabajo es encontrar modelos en base al entrenamiento de diferentes clasificadores supervisados incorporando datos de atributos tanto químicos

Special Issue in Information Technology
Research in Computing Science 57, 2012, pp. 133-140

como atmosféricos, que son considerados predictores de niveles de la calidad del aire (en particular con respecto a ozono) [2].

A continuación se muestran distintos modelos propuestos en la literatura que estudian el fenómeno de la contaminación atmosférica.

2 Trabajo Relacionado

Se han realizado diversos estudios para inferir niveles de contaminación en el campo de la estadística tradicional y para predecir en el campo de los árboles de decisión y otros clasificadores. Por ejemplo Seinfeld [2] realiza un estudio que mide las tendencias de ozono usando los valores máximos diarios para estimar tendencias del mismo.
En México existen trabajos relacionados para evaluar la contaminación en particular por ozono [3], [4] y [5], ellos intentan predecir las áreas con mayor riesgo para los habitantes de la Zona Metropolitana del Valle de México (ZMVM), mostrando tendencias de comparaciones entre algunas ciudades.
Barai [6], presenta una red neuronal para predecir la calidad del aire, ellos consideran un número limitado de datos y utilizan diferentes modelos de redes neuronales para pronosticar los valores de la calidad del aire.
Zvyagintsev [7], muestra dos métodos sintónicos estadísticos para predecir los máximos diarios que se originan en la superficie en base atributos meteorológicos haciendo uso de los datos de la capital rusa Moscú. Zvyaginsev concluye que los mejores atributos para realizar la predicción son: el tiempo, la temperatura y la humedad relativa.
Cortina-Januchs [8], usa los datos de partículas menores a 10 microgramos (PM10) de la ciudad de Salamanca, México, con el objetivo de crear un sistema de alarma para predecir las concentraciones en dicha ciudad. Usan los modelos de redes neuronales y consideran otros atributos como: dirección del viento, temperatura y humedad relativa.
Loya [9], usa procesos de Procesos de Poisson No-Homogéneos con el objetivo de modelar el fenómeno de la contaminación por ozono de la Ciudad de México, como resultado se obtienen cinco modelos diferentes para cinco estaciones meteorológicas: Pedregal, Plateros, Tlalnepantla, Merced y Xalostoc.

A continuación se explican las características de la zona que es objeto de la presente investigación.

3 Caso de estudio

La ZMVM se encuentra ubicada en una cuenca que restringe la libre circulación del viento sin permitir una buena ventilación, tiene un parque vehicular cercano a los 4 millones de vehículos, donde operan aproximadamente 30,000 industrias [10].

Para desarrollar este estudio se usa el conjunto de datos del subsistema Red de Monitoreo Automática (RAMA) y la Red de Meteorología y Radiación Solar (REDMET) considerando las mediciones horarias del periodo Enero 2010 a Diciembre de 2011, se toman en cuenta tres estaciones meteorológicas: Pedregal, Tlalnepantla y Xalostoc debido a que en estas se tienen datos confiables que han sido validados por diversos organismos.

En la Tabla 1 se muestra los atributos considerados en el estudio, junto con su abreviatura, los niveles máximos permitidos para cada contaminante según la Norma Oficial Mexicana (NOM-1993) [11] y la unidad de medida de cada atributo.

Tabla 1. Atributos químicos y atmosféricos considerados en el estudio.

Atributo químico/ atmosférico	Valor máximo permitido según NOM-1993	Unidad de Medida
Ozono (O_3)	0.11ppm	Partículas por millón
Monóxido de carbono (CO)	11 ppm	Partículas por millón
Dióxido de nitrógeno (NO_2)	0.21ppm	Partículas por millón
Dióxido de azufre (SO_2)	0.13ppm	Partículas por millón
Temperatura (TMP)		Grados Celsius
Humedad Relativa (HR)		Porcentaje (%)
Velocidad del viento (WSP)		Metros sobre segundo (m/s)
Dirección del viento (WDR)		Grados Norte

Se aplican técnicas de minería de datos con el objeto de obtener un conjunto robusto y por lo tanto susceptible de aplicar algoritmos de clasificación, haciendo uso de la herramienta Weka [12]. El preprocesamiento de los datos es mostrado en la siguiente sección.

4 Pre-procesamiento de datos

Se realiza una integración de datos en donde se mezclan diferentes conjuntos de los mismos correspondientes al SIMAT, a continuación se muestra dicha integración.

4.1 Integración de datos

Las bases de datos iniciales contienen información que no es relevante para este estudio, por consiguiente se lleva a cabo un filtrado de la información que es importante y se realiza una integración de los datos, que son útiles dado que se toman de diversas fuentes, en este caso de al menos dos bases de datos diferentes. En particular se conservan los atributos: CO, NO_2, SO_2, TMP, RH, WDR, WSP y la

HORA. Se descartan datos de otras estaciones meteorológicas y otro tipo de atributos como: partículas menores a 10 microgramos, radiación solar y precipitación fluvial. Como se usa una clasificación supervisada se estiman las clases en base a los intervalos estudiados en la NOM-1993, los cuales se muestran en la Tabla 2.

Tabla 2. Rangos de concentración para la asignación de clases.

Intervalo O_3	Clase	Clase de la calidad del aire
0.000-0.055	Verde	Buena
0.056-0.110	Amarillo	Regular
0.111-0.165	Naranja	Mala
0.166-0.220	Rojo	Muy Mala
>0.220	Morado	Altamente Mala

4.2 Limpieza de los datos y valores perdidos

Aunque es bien sabido que los clasificadores que se usan pueden trabajar con ruido y valores pérdidos, se desarrolla una estimación para aproximar estos últimos. Suponiendo que el valor perdido es V_n, se puede realizar una estimación obteniendo un promedio en base al valor V_{n-1} y V_{n+1}.

4.3 Selección de atributos

Para esta tarea se utiliza "ChiSquaredAttibuteEval" definido en el paquete Weka, el cual evalúa el valor de un atributo calculando el valor del estadístico chi-cuadrado. Los resultados son mostrados en la Tabla 3 y pueden ser interpretados de la siguiente manera: el atributo más significativo es el atributo que tenga el valor más cercano a 1, en consecuencia el atributo menos relacionado, es el más alejado de 1. Como resultado se observa que los atributos que teóricamente aportan más información y mejoran la precisión de los clasificadores son: CO, NO_2, SO_2, TMP, RH y la HORA. Por lo tanto se descarta los atributos WDR y WSP.

Tabla 3. Valores para la selección de atributos.

Atributo	PEDREGAL	TLALNEPANTLA	XALOSTOC
HORA	3.8+0.4	3.2+0.7	3.3+0.4
CO	3.1+0.5	2.9+0.9	2.0+0.0
NO_2	5.9+0.3	3.7+1.3	6.1+0.8
SO_2	2.1+0.3	4.8+1.6	6.5+2.2
TMP	1.0+0.0	1.0+0.0	1.0+0.0
RH	5.1+0.3	5.6+0.4	5.0.+0.6
WDR	7.6+0.4	7.9+0.3	7.3+0.4
WSP	7.4+0.4	6.9+0.3	4.8+0.7

A continuación, se discute como se obtiene una muestra aleatoria estratificada.

4.4 Muestra aleatoria estratificada y desbalanceo de clases

Se obtiene un muestreo estratificado de los datos, donde los estratos son las estaciones del año (primavera, verano, otoño e invierno). Para asegurar que las clases sean significativas se observan dos muestras adicionales, las cuales tienen el mismo comportamiento en lo que respecta a la cantidad de instancias correctamente clasificadas.

4.5. Clasificadores utilizados

Los clasificadores usados son los estudiados en [13]:

- *Arboles C4.5*: Se basa en la construcción del árbol de decisión a partir de un grupo de datos de entrenamiento usando el concepto de entropía de la información [14].
- *Decisicion Stump*: Es un modelo de aprendizaje automático que consiste de un árbol de decisión en un nivel, es decir un árbol de decisión con un nodo el cual esta inmediatamente conectado a sus nodos terminales u hojas. La predicción se realiza basándose en el valor de únicamente una característica [15].
- *Naïve Bayes*: Es un clasificador con asignación de probabilidades, en particular en el teorema de Bayes, se suele suponer independencia en las variables predictoras [16].
- *Random Trees*: Es un clasificador basado en la teoría de grafos, donde se tiene un grafo dirigido, en el cual para un vértice raíz U y cualquier otro vértice V, existe exactamente un único camino de U a V. Este algoritmo considera un proceso estocástico para realizar la clasificación [17].
- *Random Forest*: Agrupa diversos arboles de clasificación, para catalogar un nuevo objeto de un vector de entrada. Cada árbol proporciona una clasificación y vota por la clase a la cual pertenece el nuevo objeto. El bosque de arboles realiza la clasificación tomando en cuenta la mayoría de votos del nuevo objeto [17].
- *Multilayer Perceptron*: Es una red neuronal, tiene una capa de entrada, una o más capas ocultas y una capa de salida. Generalmente usa el algoritmo de backpropagation para su entrenamiento que es basado en el ajuste de pesos [18].

A continuación se muestra los resultados obtenidos.

5 Resultados

Inicialmente se considera un conjunto de 420,800 datos, posteriormente para cada estación meteorológica se tiene una muestra estratificada de 14,000 instancias obtenida en base a la fase de preprocesamiento, las muestras son utilizadas para realizar el entrenamiento de los modelos, para verificar se usa validación cruzada de 10 pliegues.
Los atributos seleccionados para realizar la clasificación son: CO, NO_2, SO_2, TMP, HR, HORA, y CLASE. El modelo de predicción usado está basado en las observaciones de los niveles de los atributos químicos y atmosféricos elegidos a partir de la selección de los mismos, es decir si se desea predecir la calidad del aire en la hora *n* del día. Los clasificadores deben ser capaces de: observar el comportamiento de los demás atributos en ese instante de tiempo, aproximar un valor de ozono y pronosticar en base al modelo obtenido la calidad del aire como: Buena, Regular, Mala, Muy mala y Extremadamente Mala.

En la Tabla 4 se muestra el porcentaje de instancias correctamente clasificadas (CCI, por sus siglas en Inglés) de los clasificadores estudiados. Después de probar diferentes configuraciones para cada clasificador, los parámetros con mejor nivel de precisión son los siguientes:

- C4.5: Confidence factor=0.05, MinNumObj=2, seed=1, NumFolds=3, Poda=Falso
- DecisionStump y Naive Bayes: Valores por default.
- RandomTree: K-value= 0 MinObjects =1.0 , Seed= 1
- RandomForest: -Num trees=50 -Num features= 0 -S 1 -depth 10 -num-slots 1
- MultilayerPerceptron: 1 Layer, 500 epochs training y 8 neurons by layer.

Tabla 4. Niveles de precisión globales por clasificador, en la primera columna se tiene el clasificador usado, en la segunda, tercera y cuarta se muestra en porcentaje, la cantidad de CCI para cada estación meteorológica estudiada. Se muestra mediante el sombreado el clasificador con el cual se obtuvieron los mejores resultados.

Clasificador	PED	TLA	XAL
C4.5	91.20	82.49	91.40
Decision Stump	73.91	62.36	81.44
Random Tree	88.30	80.86	89.45
Random Forest	92.14	85.37	92.23
Naïve Bayes	85.83	74.40	82.37
Multilayer Perceptron	88.71	84.94	88.28

Se puede observar que el mejor modelo para las tres estaciones meteorológicas está dado por el clasificador basado en árboles *Random Forest*. Para todas las estaciones meteorológicas se observa que la mayoría de los clasificadores supervisados muestran una buena precisión al categorizar los niveles de contaminación.

Por otra parte en conjunto, los arboles de decisión C4.5 y *Random Forest* presentan un mejor comportamiento que los clasificadores basados en probabilidad como *Naïve Bayes* y redes neuronales como Multilayer Perceptron.

En la Tabla 4 se puede observar que la precisión para cada estación meteorológica está en el intervalo [73.91-92.14] para el caso de Pedregal, [62.36-85.37] Tlalnepantla y [81.44 – 92.23] para Xalostoc. Se observa que la clase con mejor precisión es "Verde" o "Buena" debido a que los datos de entrenamiento en esa clase son claramente identificables a diferencia del resto de clases en las cuales los valores límite son más estrechos.

6. Conclusiones y trabajo futuro

La implementación realizada para estimar la calidad del aire con respecto a ozono, puede ser usada como herramienta para la toma de decisiones, planeación y evaluación de la calidad del aire, inclusive para informar a las personas acerca de alguna emergencia. En este documento son mostrados modelos para el pronostico de la calidad del aire basados en el entrenamiento de clasificadores supervisados, considerando los datos de atributos químicos y atmosféricos del SIMAT de la Ciudad de México.

Este trabajo será presentado a las autoridades del SIMAT, ya que puede ser de gran ayuda en la prevención de diversos fenómenos que se originan a raíz de la contaminación ambiental subyacente en la Ciudad de México.

Finalmente como trabajo futuro se pretende investigar acerca de otras variables químicas que pueden influir substancialmente en los niveles de ozono, por ejemplo partículas menores a 10 microgramos, precipitación fluvial, inclusive es necesario integrar otros factores como las emisiones de ceniza volcánica del Popocatépetl ubicado en las cercanías de la Ciudad de México. También se pretende extender la clasificación considerando un mayor número de estaciones meteorológicas, así como la de otros estados de la República Mexicana.

Referencias

1. Reyes, H., Vaquera, H., or, J.V.: Estimate of tendencies in high levels of urban ozone using the quantiles of the distribution of generalized extreme values 21, 470-481. (2010).
2. Seinfeld, J.: Comittee on tropospheric ozone formation and measurement; Board on Environmente Studies and Toxicology; Board on Atmospheric Sciences and Climate;

Commission on Geosciencies, Environment and Resources; National Re-search Council, Rethinking the on ozone problem in urban and regional air Pollution. National Academic Press, Washington (1991).
3. Molina, M., L., M.: The impacts of magacities on air pollution, environmental aspects of urbanization, Goteborg Sweden (2004).
4. Aguirre E., A.A.L.B.: A system for forecast of the maximun ozone levels. Atmospheric Environment 38 4689-4699, (2004).
5. I.N.E.G.I.: Estadísticas del Medio Ambiente. Semarnap, México (1999).
6. Barai, S., Dikshit, A., Sharma, S.: Neural network models for air quality prediction: A comparative study. In Saad, A., Dahal, K., Sarfraz, M., Roy, R., eds.: Soft Computing in Industrial Applications. Volume 39 of Advances in Soft Computing. Springer Berlin / Heidelberg 290- 305, (2007).
7. Zvyagintsev A.: Statistical forecast of surface ozone concentration in Moscow, Russian Meteorology and Hydrology, Allerton Press, Inc. distributed exclusively by Springer Science+Business Media LLC, 499-506, (2008).
8. Cortina-Januchs M.G., Barrón-Adame J.M., Vega-Corona A., Andina D.: Pollution Alarm System in Mexico, Proceeding IWANN '09 Proceedings of the 10th International Work-Conference on Artificial Neural Networks: Part I: Bio-Inspired Systems: Computational and Ambient Intelligence, 1336-1343, Springer-Verlag Berlin, Heidelberg (2009).
9. Loya N.: Modelación de fenómenos atmosféricos usando procesos de Poisson no-homogéneos. Tesis de Licenciatura, Benemérita Universidad Autónoma de Puebla, (2008).
10. Gobierno del Distrito Federal, Sistema de monitoreo atmosférico, Online Acceso 15 de enero de 2012. http://www.sma.df.gob.mx/simat/
11. Norma Oficial Mexicana (NOM-1993), disponible en la página oficial del SMA: http://www.sma.df.gob.mx/sma/links/download/archivos/anterior_NOM-020-SSA1-1993.pdf
12. Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P., Witten I.H.: The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1. (2009).
13. Mitchell, T.M.: Machine Learning. 1 edn. McGraw-Hill, Inc., New York, NY, USA (1997).
14. Quinlan J.: Programs for machine learning, Morgan Kaufmann Publishers, (1993).
15. Iba Ai A., Langley P.: Induction of One-Level Decision Trees, Proceedings of the Ninth International Conference on Machine Learning, Morgan Kaufmann, 233-240 (1992).
16. John G., Langley P.: Estimating Continuous Distributions in Bayesian Classifiers. In: Eleventh Conference on Uncertainty in Artificial Intelligence, San Mateo, 338-345, 1995.
17. Breiman L.: Random Forests. Machine Learning. 5-32. (2001).
18. McCulloch W., Pitts W.: A logical calculus of the ideas immanent in nervous activity, Bulletin of Mathematical Biology, Vol. 5, No. 4 115-133, (1993).