Red Neuronal Auto-Organizada con Aprendizaje en Tiempo Real para la Predicción de la Calidad del Aire en base a PM_{10} en Villahermosa Tabasco, México

Jesús Carrera-Velueta[1], Elizabeth Magaña-Villegas[1], Carlos Gónzalez-Figueredo[2], José Hernández-Barajas[1], Sergio Ramos-Herrera[1], Raúl Bautista-Margulis[1], José Laines-Canepa[1], Arturo Valdez-Manzanilla[1].

[1] Universidad Juárez Autónoma de Tabasco, División Académica de Ciencias Biológicas, Carr. Vhsa-Cárdenas entronque Bosques de Saloya. 86139. Villahermosa, Tabasco. México. {jesus.carrera, elizabeth.magana, roberto.hernandez, sergio.ramos, raul.bautista, jose.laines, arturo.valdez}@ujat.mx

[2] Instituto Tecnológico y de Estudios Superiores de Occidente, Departamento de Procesos Industriales y Tecnológicos, Av. Periférico Sur Manuel Gómez Morín 8585, Tlaquepaque Jal., 45604, figueredo78@hotmail.com

Resumen. Se diseño un modelo de red neuronal artificial para la predicción al día siguiente del máximo diario de PM_{10}, (material particulado de menos de 10 micrometros de diámetro), el cual se construye de manera dinámica mediante la formación de *clusters* para la clasificación de patrones y evoluciona a través de los datos que recibe automáticamente y en tiempo real. Se generó una matriz de distancias a partir de los patrones de entrada para seleccionar el radio óptimo de clasificación. El modelo fue validado mediante la aplicación de datos históricos de variables meteorológicas y de PM_{10} registrados en Villahermosa, Tabasco, México de 2007 a 2009. Los experimentos realizados permitieron identificar las variables relevantes del modelo y se contemplaron datos normalizados y no-normalizados. Los mejores resultados del modelo se obtuvieron usando promedios móviles y valores máximos y mínimos de PM_{10} no normalizados como variables de entrada así como radios cercanos al valor mínimo calculado en la matriz de distancias.

Palabras claves: Redes Neuronales Artificiales, Predicción de la contaminación del aire, Material Particulado, PM_{10}.

1 Introducción

El uso de las Redes Neuronales Artificiales (RNA) para la predicción de los promedios diarios en el nivel de concentración de material particulado (PM) es una práctica generalizada en diferentes países, dados los resultados que éstas tienen superiores a los obtenidos por métodos estadísticos tradicionales. Corani en Italia [1], Jef en Bélgica [2], Chiarvetto en Argentina [3], Pérez y Reyes en Chile [4],[5], son algunos de los autores que han elegido como herramientas para sus predicciones de promedios diarios para diferentes contaminantes, el uso de RNA. Los mejores resultados reportados fueron usando una red *back-propagation*, sin embargo, requieren mucho tiempo de cómputo para sus entrenamientos y un enorme número de experimentos para encontrar lo parámetros idóneos para una predicción exitosa, como son el coeficiente de aprendizaje y el momento.

Baeza desarrolló un modelo predictivo de los niveles máximos de PM_{10} usando una RNA de tipo *back-propagation*. El estudio se realizó con información meteorológica y de PM_{10} en el periodo de 2007 a 2009 obteniéndose un porcentaje de predicción del 50% de los datos validados [6]. El objetivo de la presente investigación fue diseñar e implementar una Red Neuronal Artificial para la Clasificación, Agrupación y Asociación de patrones (CLASO) basada en la formación de *clusters* en tiempo real para el análisis de patrones para pronóstico de la calidad del aire en base a PM_{10} en la atmósfera en Villahermosa Tabasco. Para el desarrollo de la red CLASO, se tomó como base el trabajo desarrollado por Márquez [7] en el que se propone un modelo de red neuronal para la clasificación y la agrupación de patrones genómicos. Esta red se construye de manera dinámica y trabaja en dos etapas, está basada en el aprendizaje competitivo y fusiona las fases de entrenamiento y validación para tener un aprendizaje en tiempo real.

2 Metodología del Modelo CLASO y su aplicación en la predicción de contaminantes de PM10.

Se conformaron bases de datos que incluyeron información horaria de las variables meteorológicas (temperatura, humedad relativa, radiación solar, velocidad y dirección del viento), variables estacionales (año, mes y día) y de $PM_{10.}$ de los años 2007 a 2009. Los patrones de entrada se conformaron por variables meteorológicas con valores máximos diarios y variables de PM_{10} correspondientes a promedios móviles, valores máximos y mínimos y promedios diarios de los datos horarios que registra la estación de monitoreo, por lo que cada patrón presentado a la red considera las condiciones de Calidad del Aire de por lo menos 48 horas previas a la de predicción.

El modelo aprende a clasificar patrones a partir de una base de datos históricos y realiza esta tarea de un modo supervisado usando como criterio las distancias entre ellos. La estructura de la red queda constituida por la capa de entrada, una capa oculta formada por centroides organizados en clusters o cúmulos y una capa de salida, cuyas categorías están representadas por los niveles máximos de concentración de PM_{10}. Una vez que la topología de la red se construye, el funcionamiento de la red cambia para agrupar nuevos patrones y para predecir los valores máximos de PM_{10} del día siguiente.

Para iniciar el funcionamiento del modelo, se calculan las distancias entre los patrones de entrada usando la medida de distancia Euclidiana. Este proceso genera una matriz de distancias que nos permite conocer la distancia máxima y mínima que hay entre ellos. De estos valores se selecciona el valor mínimo como radio inicial de los centroides. En principio la red sólo cuenta con la capa de entrada, no tiene capa oculta ni capa de salida, aunque se conocen los niveles máximos de contaminación (clases o categorías para la clasificación de los patrones), éstos no se asignan hasta que se presenten los patrones a la red. Con el primer patrón que se presenta a la red se crea el primer centroide de la capa oculta asignándole como radio inicial el valor obtenido en la matriz de distancias y el cual representa su región de clasificación. Como podemos observar en la Figura 1, los valores de este primer patrón se convierten en el vector de pesos para dicho nodo, el cual también conformará el primer grupo (*cluster*) y se le asignará a éste la clase a la que pertenece, registrada en la base de datos históricos.

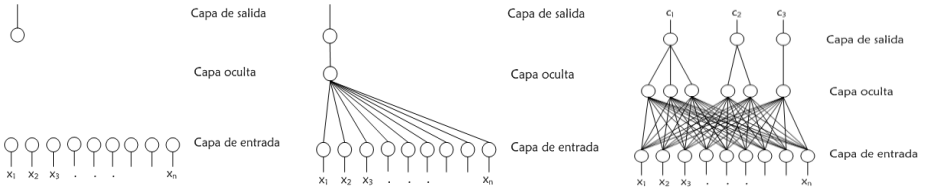

Figura. 1. Estado inicial de la red antes de presentar el primer patrón, después de crear el primer centroide y estado final de la topología de la red.

Con la presentación del segundo patrón, se calcula la distancia entre el vector de entrada y el vector de pesos, del primer patrón (en este caso, el primer módulo oculto creado) para determinar si es clasificado dentro de este primer grupo, en caso contrario, se crea un nuevo centroide para la capa oculta. De manera sucesiva, este proceso se repite hasta que se presentan todos los patrones a la red. De este modo se construye la capa oculta que finalmente tendrá un conjunto de neuronas organizadas en un número determinado de *clusters* o cúmulos, los cuales estarán representados en la capa de salida por una clase conocida (Figura 1). Cada una de las neuronas o centroides agrupados en los *clusters* habrán clasificado uno o más patrones de la base de datos. Una vez presentados todos los patrones a la red, se consideran los centroides que al final de una época clasificaron un solo patrón, formándose así una nueva base de datos que es presentada nuevamente a la red para su reclasificación. Por lo que la red se construye con la presentación del conjunto de patrones en más de una ocasión.

El siguiente paso a la construcción del modelo es su actualización permanente, a través de la alimentación de nuevos patrones que recibe en tiempo real y de manera automática, de la estación de monitoreo a la que esté conectada. Trabaja de un modo no supervisado en el que tiene la tarea de agrupar nuevos patrones. En este proceso, la agrupación de los patrones se realiza a partir del vector de pesos generados en el módulo de clasificación. Si un nuevo patrón es presentado a la red y no puede ser clasificado por ninguno de los centroides existentes, se crea un nuevo centroide y se calcula el valor de una nueva clase, la cual es asignada mediante el cálculo de las distancias entre el patrón de entrada y los dos centroides más cercanos. El cálculo se realiza como la proporción entre la distancia más corta y la distancia total entre los centroides clasificadores. Por último, el proceso de predicción consiste en presentar un nuevo patrón a la red para agrupar el nuevo patrón dentro de los centroides creados. Si el patrón es ubicado en alguno de los *clusters* formados se realizará la predicción como una asociación, mediante una función lineal, con la clase del centroide ganador.

3 Resultados y Discusión

La combinación de variables también fue un elemento importante para el funcionamiento eficiente del modelo. Se observó que la red trabaja mejor cuando utiliza solo valores de concentración de PM_{10}. El conjunto de patrones correspondientes al año 2009 fueron los que mostraron mejores resultados. En el proceso de validación del modelo, la predicción de los patrones no normalizados mostró un 72% de éxito de los datos procesados con un error cuadrático medio de 8.2 $\mu g/m^3$. Para este conjunto de patrones, en todos los experimentos, se observaron 5

patrones, que de manera recurrente presentaron valores de predicción muy alejados del valor real, por lo que se presume tienen un error en el monitoreo. Para el análisis de los resultados, estos patrones fueron eliminados, consecuentemente se observó que el porcentaje de predicción aumento a un 80% y el error cuadrático medio bajo a 4.5 $\mu g/m^3$. Tomando en consideración que la distancia entre las clases es de 3 $\mu g/m^3$, un error cuadrático medio de 4 $\mu g/m^3$ se considera aceptable. Además se estimó el porcentaje de predicciones cuya distancia de predicción fuera menor y/o igual a $3\mu g/m^3$, mostrando un porcentaje predictivo del 71.4%. Tomando en cuenta que en las investigaciones documentadas se manejan porcentajes de predicción de hasta el 67%, los resultados obtenidos con el CLASO se pueden considerar favorables.

4 Conclusiones

El modelo propuesto no requiere de excesivos tiempos de procesamiento para su aprendizaje. No se necesita de cantidades grandes de patrones para su construcción y su funcionamiento y rendimiento están en función de la base de datos históricos con la que es alimentada en su fase inicial de clasificación. Es sensible al radio inicial de clasificación, siendo los valores más adecuados los más cercanos a la distancia más corta entre ellos, obtenida de la matriz de distancias. La generación de ésta matriz al iniciar la ejecución del modelo resulta entonces esencial para definir este parámetro. Se mejoró los resultados reportados por la red *back-propagation* en el trabajo de Baeza en 2010, misma que fue aplicada a la predicción de material particulado de menos de 10 micrómetros. Por otro lado, de acuerdo los resultados reportados por otros autores, los mejores porcentajes de predicción son del 67%, en trabajos realizados usando redes neuronales, principalmente en Chile, por lo que podemos concluir que los resultados del modelo CLASO son satisfactorios.

Referencias

1. Corani, G. Air quality prediction in Milan: feed-forward neuronal networks, pruned neuronal networks and lazy learning. Ecological Modelling, 513-529, 2005. ISSN 0304-3800 (2005).
2. Jef et. al. A neural network forecast for daily average PM_{10} concentrations in Belgium. Atmospheric Enviroment. 39: 3279-3289. ISSN 1352-2310. (2005).
3. Chiarvetto Peralta et al. Aplicación de redes neuronales artificiales para la predicción de calidad de aire. Mecánica Computacional Vol. XXVII. (2008).
4. Pérez y Reyes. Prediction of $PM_{2.5}$ concentrations several hours in advance using neural networks in Santiago, Chile. Atmospheric Environment. ISSN 1352-2310. (1999).
5. Pérez y Reyes. Prediction of maximum of 24-h average of PM_{10} concentrations 30h in advance in Santiago, Chile. Atmospheric Environment. ISSN 1352-2310. (2002).
6. Baeza M. S. Predicción de la calidad del aire en base a PM_{10} en Villahermosa, Tabasco mediante la aplicación de Redes Neuronales Artificiales. Tesis de Licenciatura. Universidad Juárez Autónoma de Tabasco. (2010).
7. Márquez M. C. E. Análisis de patrones de expresión de genes utilizando redes neuronales (por clustering). Tesis de Maestría. Universidad Nacional Autónoma de México. (2004).