ALGORITMOS DE APRENDIZAJE AUTOMÁTICO: APLICACIÓN EN LA SOLUCIÓN A PROBLEMAS MEDIOAMBIENTALES

Anabel Vega Calcines*

Resumen

El Aprendizaje Automático es una rama de la Inteligencia Artificial que tiene por objetivo desarrollar técnicas mediante las cuales las computadoras puedan aprender y ayudar al hombre en la solución de problemas complejos. Los algoritmos de aprendizaje pueden clasificarse de acuerdo a la cantidad de datos de entrenamiento disponibles, en algoritmos supervisados, no supervisados y semisupervisados. La aplicación de uno u otro algoritmo depende de las características del asunto a resolver, pues cada uno de ellos puede ser útil en determinados circunstancias. En el presente trabajo se describen brevemente los algoritmos antes mencionados y se muestran sus potencialidades para resolver problemas medioambientales a través de ejemplos prácticos.

Palabras claves: *Aprendizaje automático, algoritmos supervisados, algoritmos no supervisados, algoritmos semisupervisados, problemas medioambientales.*

---

* Ingeniera en Ciencias Informáticas, Universidad de Ciencias Informáticas. Cuba. Estudios de Maestría en Informática para la Gestión Medioambiental, Universidad Central “Martha Abreu” de Las Villas. Correo electrónico: anabelv@uclv.edu.cu

Introducción

El Aprendizaje Automático es una rama de la Inteligencia Artificial que tiene por objetivo desarrollar técnicas mediante las cuales las computadoras puedan aprender a desarrollar tareas que los seres humanos hacemos de forma natural y rápida, como por ejemplo, reconocer imágenes, entender el lenguaje natural, tomar decisiones, etc.

Los algoritmos de aprendizaje se pueden clasificar a través de distintas dimensiones, de la representación de conocimiento utilizada y de las fuentes a partir de las cuales se obtienen las experiencias de aprendizaje (ej. Datasets, Entornos de simulación). Una de las dimensiones más importantes en el área de Aprendizaje Automática es la cantidad de datos de entrenamiento disponible para el sistema. Los datos de entrenamiento son pares de entradas y salidas deseadas. En esta dimensión aparece el aprendizaje supervisado (gran cantidad de datos de entrenamiento) y no supervisado (ausencia total de datos de entrenamiento). Entre estas dos clasificaciones está el aprendizaje semisupervisado. (Damirechi and López, 2009)

Las principales técnicas para desarrollar el Aprendizaje Automático son: las redes neuronales, el aprendizaje inductivo y el razonamiento basado en casos (CBR). La diferencia fundamental entre estas técnicas radica en la forma en que se almacena el conocimiento. Así, en las redes neuronales, el conocimiento se traduce en una serie de pesos y umbrales que poseen las neuronas. En cambio, en el aprendizaje inductivo, el conocimiento se transforma en un árbol de decisión o un conjunto de reglas. Por último, en el CBR, el conocimiento está formado por una base de casos compuesta por los problemas resueltos en el pasado. (Priore et al., 2002)

El Aprendizaje Automático tiene una amplia gama de aplicaciones. El presente trabajo tiene por objetivo exponer, a través de ejemplos, los resultados de la aplicación de algoritmos supervisados, no supervisados y semisupervisados en la solución a problemas medioambientales.

1 Aplicación de algoritmos supervisados en la solución de problemas medioambientales

El aprendizaje supervisado consiste en entrenar un sistema partir de un conjunto de datos etiquetados o patrones de entrenamiento, compuesto por patrones de entrada y la salida deseada. El objetivo del algoritmo es desarrollar una función capaz de deducir el valor correspondiente a cualquier entrada válida, de manera tal que la salida generada sea lo más cercanamente posible a la verdadera salida dada una cierta entrada. El patrón de salida hace el papel de supervisor. (Cortina, 2012)

Usualmente, se poseen grandes cantidades de datos históricos con información medioambiental de diversos ámbitos, que pueden servir para predecir situaciones similares. Estos datos pueden ser útiles como patrones de entrenamiento, en la implementación de algoritmos de aprendizaje supervisado. Se han obtenido muy buenos resultados en tareas de clasificación y de predicción.

A lo largo de los años diferentes modelos de Redes Neuronales Artificiales (RNA), se han aplicado por diversos autores en la predicción concentraciones promedios de contaminantes atmosféricos. Algunos modelos utilizan datos de contaminación pasada, más información meteorológica y otros datos útiles. A continuación se exponen algunos ejemplos:(Cortina, 2012)

En la ciudad de Palermo, Italia se desarrolló un sistema que predicen la máxima concentración para los contaminantes SO_2, O_3, PM10, NO_2 y CO utilizando una base de datos de tres casetas de monitorización ambiental. El modelo propuesto es una red recurrente. Las variables utilizadas para la predicción fueron: dirección del viento, velocidad del viento, presión atmosférica y temperatura ambiente. La topología de la red fue de 9 neuronas de entrada y una neurona en la capa de salida. Los parámetros fueron evaluados mediante la Media del Error Absoluto, la Raíz del Error Cuadrático Medio, el Error Cuadrático Medio y el Coeficiente de Correlación, obteniendo un Coeficiente de Correlación en el rango de 0.72 a 0.97 para cada contaminante.

El modelo Box-Jenkins (modelo autoregresivo de media móvil) univariante fue implementado para la predicción de la concentraciones de SO_2, NO_2 y Partículas Suspendidas Totales (PST) la Ciudad de Delhi obteniendo un *d* (índice de acuerdo, congruencia entre los datos predichos y observados) entre 0.88 a 0.91 para el SO_2 y el NO_2, y 0.80 a 0.91 para las PST. El único requisito es la disponibilidad de suficientes datos históricos para la formulación del modelo, indican que si es posible, por los menos 50 y preferiblemente 100 observaciones sucesivas deben estar disponibles para hacer la predicción.

(Silva et al., 2003) realizó una evaluación de cuatro métodos para la predicción de contaminantes atmosféricos (PM2,5 y PM10) en Santiago de Chile, utilizando como variables de entrada: la temperatura, humedad relativa, velocidad del viento, dirección del viento, mes del año, día de la semana, hora de registro y nivel de concentración de material particulado (se utilizaron 3754 observaciones). El estudio fue realizado a través de discriminantes no paramétricos, un modelo de regresión lineal múltiple, una RNA multicapa de propagación hacia atrás y modelos MARS (Multivariate Adaptive Regression Splines). La metodología MARS se centra en la construcción de un modelo de regresión no-lineal, basado en un producto de funciones base spline. El modelo de respuesta construido, automáticamente selecciona las variables predictoras y detecta posibles interacciones entre ellas generando modelos más flexibles. Estas interacciones quedan

expresadas algebraicamente a través de las funciones basales. Los resultados arrojados por las diferentes metodologías se compararon en varias estaciones de medición, a pesar de que todas las metodologías mencionadas ofrecen modelos adecuados para estudiar la contaminación por material particulado, MARS resultó ser superior en cuanto a exactitud y rapidez.

La efectividad de la aplicación de RNA también ha sido demostrada en la modelación de procesos de tratamiento de aguas residuales. Se han usado redes de propagación hacia adelante ("feedforward"), con algoritmos de aprendizajes de retropropagación ("backpropagation") con este propósito, que han arrojado resultados superiores a los obtenidos por modelos de regresión múltiple. También se han diseñado redes para la predicción de la Demanda Biológica de Oxígeno (DBO) en el efluente de salida, usando redes de perceptrón multicapa. (Rodríguez et al., 2005)

Los sistemas de CBR igualmente han sido empleados para el diseño de operaciones más apropiadas, de acuerdo a un conjunto determinado de contaminantes de entrada en las plantas de tratamiento de aguas residuales. (Rodríguez et al., 2005)

En materia forestal, se han desarrollado varios algoritmos de aprendizaje supervisado que abordan tareas de clasificación y predicción como los que siguen.

Un grupo de investigadores de la División de Ciencias Forestales de la Universidad Autónoma Chapingo, México, implementaron un método para la obtención de tipos de coberturas forestales mediante RNA. Las RNA desarrolladas en este trabajo se basaron en información geográfica (altitud, exposición, pendiente, distancia a los escurrimientos, geología y edafología) e imágenes de satélite haciendo uso del análisis de componentes principales, para definir la variable dependiente (vegetación). Esta información fue procesada con una RNA de retropropagación con dos capas ocultas, con sus respectivas funciones de activación (tangencial hiperbólica y gaussiana). Obteniendo un error cuadrático medio de 0.8617 para la fase de entrenamiento y 0.8514 en la fase de prueba, alcanzando un 83 % de sitios predichos correctamente, sobrepasando lo alcanzado por otros autores con métodos tradicionales. (Rodríguez et al., 2002a)

La mortalidad regular de recursos forestales, es aquella producida por factores no catastróficos. Guan y Gertner, 1991 diseñaron una red neuronal con algoritmo Backpropagation, dos variables dependientes del modelo logístico y 11 unidades ocultas. Además prueban 5, 7, 10 y 20 unidades en la capa oculta y 7 unidades en la capa de entrada. Las unidades en la capa de entrada son el resultado de combinar las variables independientes en lo que se llaman "patrones de actividades" basados en la distribución Gaussiana utilizando cuatro clases diamétricas y tres clases de incremento. La mejor estimación de la mortalidad se obtuvo con la red de 5 unidades ocultas y 7 inputs. (Bravo-Oviedo and Kindermann, 2004)

Por otra parte, la modelización de los procesos ecológicos, resulta de gran interés para los investigadores de esa rama.(Bravo-Oviedo and Kindermann, 2004) describe un algoritmo que determina la densidad de alevines de trucha que se obtiene de la reproducción de acuerdo a las características del hábitat. El algoritmo consta de una red de tres capas que implementa un algoritmo de retropropagación. La capa de entrada constaba de 10 unidades inputs referidas al hábitat de la trucha en el estudio de su reproducción, 8 unidades ocultas en la capa oculta determinadas de manera empírica y una única unidad de salida output que representaba la densidad de los alevines de trucha por medio lineal de corriente. La red fue comparada con un modelo de regresión múltiple lineal y encontraron mayor precisión con el modelo neuronal (error cuadrático medio de 0.785 para el modelo de la red frente a 0.371 en la regresión múltiple con datos de validación).

La modelización diaria de las aportaciones de lluvia a los ríos, es decisiva para la correcta planificación de la generación en sistemas hidroeléctricos. En(Méndez, 2009), se propone un modelo de predicción con RNA, que busca solucionar las deficiencias del modelo clásico de series temporales Box-Jenkins, factible a largo y mediano plazo, pero con un perfil bastante errático a corto plazo. El modelo de RNA propuesto es un perceptrón multicapa con una capa oculta. El algoritmo de entrenamiento que emplea es retropropagación. La variable que estudia y predice es una variable continua, por lo que se eligió el error cuadrático medio para la función de error de la red. Los resultados obtenidos fueron favorables, pero se destaca la importancia de tener un conjunto amplio de datos de entrenamiento para obtener predicciones más certeras y la necesidad de realizar reentrenamientos periódicos con nuevos datos, que considere la nueva información hidrográfica disponible.

2 Aplicación de algoritmos no supervisados en la solución de problemas medioambientales

En el aprendizaje no supervisado, muchas veces llamado de auto-organización, el propio sistema trata de identificar algún tipo de regularidad en un conjunto de datos de entrada sin tener conocimiento a priori, solamente requiere de vectores de entrada para adiestrar el sistema. Esto se logra mediante el algoritmo de entrenamiento, que extrae regularidades estadísticas desde el conjunto de entrenamiento. (Tello, 2006)

Una de las técnicas de aprendizaje no supervisado más aplicadas a la solución de problemas medioambientales es el *clustering* o agrupamiento. Esta técnica divide un conjunto de datos (u objetos) en una serie de subclases significativas llamadas grupos (clústeres).Un buen método de agrupamiento produce grupos de alta calidad en los cuales la similitud dentro del grupo es alta y la similitud entre las clases es baja. La medida de similitud se define usualmente por proximidad en un espacio multidimensional.(Tello, 2006)

El clustering juega un papel muy importante en aplicaciones de minería de datos, tales como exploración de datos científicos, recuperación de la información y aplicaciones sobre bases de datos espaciales (tales como SIG o datos procedentes de la astronomía).

Un claro ejemplo de sistema con aprendizaje no supervisado lo constituyen los sistemas topográficos. Estos sistemas se caracterizan por tener como entrada una Fotografía aérea y como salida la elaboración de un mapa. Esto se consigue analizando los patrones y modificando el mapa por su equivalente simbólico en el caso de los patrones más sencillos. Entre los problemas que se pueden plantear con estos algoritmos están el reconocimiento de puntos geodésicos y puestos hidrológicos, la representación de patrones cartográficos lineales y el reconocimiento de isolíneas.(Tello, 2006)

En otras investigaciones, se necesita reconocer patrones a partir de imágenes obtenidas desde satélites pues las mismas ofrecen una perspectiva única de la superficie terrestre, aportando información adicional que pasa desapercibida a los sentidos humanos.

(Rodríguez et al., 2002b) describe una solución para realizar un inventario de cultivos de la vid en Bierzo, España, utilizando la teledetección. En este caso, se realizó la clasificación de una imagen utilizando varios algoritmos. Inicialmente se utilizó una clasificación no supervisada con el algoritmo isodata. Se obtuvieron un total de 35 clases espectrales, que sirvieron como base para definir las clases informacionales y espectrales representativas. Esta clasificación se mejoró con otras de tipo supervisado y calculando índices de vegetación.

Otro ejemplo en el que se utilizan técnicas de teledetección, se desarrolló en la Universidad de Córdoba, España, donde se diseñó un algoritmo de aprendizaje evolutivo y estadístico para la determinación de mapas de malas hierbas. Este algoritmo aborda la resolución de problemas de clasificación binaria utilizando una metodología híbrida que combina la regresión logística y modelos evolutivos de redes neuronales de unidades producto. Para estimar los coeficientes del modelo lo hace en dos etapas, primero aprende los exponentes de las funciones unidades producto, entrenando los modelos de redes neuronales mediante computación evolutiva y una vez estimados el número de funciones potenciales y los exponentes de estas funciones, se aplica el método de máxima verosimilitud al espacio de características formado por las covariables iniciales junto con las nuevas funciones de base obtenidas al entrenar los modelos de unidades producto. Esta metodología híbrida en el diseño del modelo y en la estimación de los coeficientes se aplica a un problema real agronómico de predicción de presencia de la mala hierba Ridolfiasegetum Moris en campos de cosecha de girasol. Los resultados obtenidos con este modelo mejoran los conseguidos con una regresión logística estándar en cuanto a porcentaje de patrones bien clasificados sobre el conjunto de generalización. (Gutiérrez et al., 2007)

El papel de la minería de datos en el análisis sismológico es cada vez más destacado debido a que los avances en las tecnologías de captura de datos proporcionan un número inmenso de estos. (Benítez, 2010)

En Irán se utilizó un algoritmo de clustering difuso no supervisado para detectar patrones espaciales en el catálogo de sismos histórico e instrumental lo cual reveló patrones ocultos que permiten clasificar de manera diferente los epicentros espacialmente distribuidos de los terremotos. La comparación entre los resultados de este análisis y las provincias sismotectónicas de Irán mostraron que las provincias encontradas confirmaron las provincias conocidas y evidenciaron rasgos nuevos para la interpretación sismológica y por consiguiente para la evaluación de la amenaza sísmica.(Benítez, 2010)

El reconocimiento de patrones espaciales sísmicos por análisis clúster también fue propuesto como una solución para la identificación y clasificación de las provincias sismotectónicas del occidente colombiano, en combinación con el análisis exploratorio de los datos del catálogo de sismos. Los insumos básicos de esta metodología son los mapas geológicos y de fallas sismogénicas del Suroccidente Colombiano y el catálogo de sismos.(Benítez, 2010)

3 Aplicación de algoritmos semisupervisados en la solución de problemas medioambientales

La adquisición de datos etiquetados para resolver un problema suele requerir un agente humano capacitado para clasificar de forma manual los ejemplos de entrenamiento. El coste asociado al proceso de etiquetado puede hacer que un conjunto de entrenamiento totalmente etiquetado sea inviable, mientras que la adquisición de datos sin etiquetar es relativamente poco costosa. En estos casos, el aprendizaje semi-supervisado puede ser muy útil para mejorar la exactitud del aprendizaje, al combinar mediante diferentes técnicas (por ejemplo, *co-entrenamiento*) un grupo de datos etiquetados con un conjunto mayor de datos no etiquetados. (Zhu, 2008)

Una importante aplicación de los algoritmos de aprendizaje semisupervisado la observamos en la actualización de los Sistemas de Información Geográfica (SIG). Los problemas de la actualización de los SIG son el coste económico y de tiempo que requieren, ya que tradicionalmente se han realizado con técnicas de fotointerpretación y visitas al terreno.(Izquierdo et al., 2008) propone una metodología para la actualización del SIG citrícola de la Comunidad Valenciana (España) mediante técnicas automáticas a partir de ortoimágenes aéreas de alta resolución.

La metodología propuesta realiza un análisis orientado a objetos que define los recintos catastrales como entidades individuales, extrayéndose las características principales de cada parcela y clasificándola posteriormente combinando árboles de decisión, máquinas de

vectores soporte y redes neuronales artificiales. El conjunto de entrenamiento se obtuvo mediante el etiquetado de parcelas por fotointerpretadores expertos. Se seleccionaron parcelas de diversos municipios cubriendo todas las tipologías de cultivo. Con todas estas muestras etiquetadas se entrenaron tres clasificadores globales para cada provincia y tres más para cada tipología de cultivo de cada provincia. Se realizó una combinación lineal entre el resultado de los distintos clasificadores y la clasificación del SIG citrícola anterior. Con la combinación lineal se obtuvo la clase final de las parcelas y un grado de fiabilidad definido por la discrepancia entre los distintos clasificadores y la clase del SIG anterior. Por medio de la clasificación automática se pudo clasificar el 87% de las parcelas procesadas de la Comunidad con un acierto superior al 92% en las tres provincias.

Conclusiones

Los problemas medioambientales por lo general son complejos y manejan conjuntos de datos crecientes. Los algoritmos de aprendizaje automático son herramientas muy eficaces en la solución de este tipo de problemas, por su capacidad de mejorar dinámicamente en la medida que se van entrenando. La elección del algoritmo más apropiado depende de las características del problema en cuestión.

Los algoritmos de aprendizajes supervisados han demostrado ser muy eficaces en la solución de aquellos problemas medioambientales en los que se posee un conjunto amplio de datos históricos de entrada con las salidas correspondientes, siendo más certeros los resultados en la medida que es más extenso el conjunto de patrones de entrenamiento.Han arrojado resultados favorables en tareas de predicción y clasificación. El paradigma de las RNA feedforward con entrenamiento de retropropagación ha demostrado su efectividad en este tipo de problemas.

Los algoritmos no supervisados, implementados principalmente a través de las técnicas de clustering,son útiles en el reconocimiento de patrones, pues a partir fotografías aéreas, imágenes de satélites, mapas y catálogos, logran extraer información no trivial sin tener conocimiento a priori. Con estas técnicas se pueden obtener resultados favorables en el campo de la topografía, análisis de cultivos y terrenos, entre otros.

Los algoritmos semisupervisados pueden ser factibles cuando no es posible realizar un amplio etiquetado de los datos y se necesita aplicar otras técnicas para mejorar la exactitud del aprendizaje, siendo provechosas en la actualización automática de SIG.

Referencias

1. BENÍTEZ, H. D. 2010. Reconocimiento de patrones espaciales sísmicos en el suroccidente colombiano.
2. BRAVO-OVIEDO, A. & KINDERMANN, G. 2004. Modelización de sistemas ecológicos mediante redes neuronales. *Cuad. Soc. Esp. Cienc. For.* , 18**,** 311-316.
3. CORTINA, M. G. 2012. *Aplicación de técnicas de Inteligencia Artificial a la predicción de contaminantes atmosféricos.* Tesis Doctoral, Universidad Politécnica de Madrid.
4. DAMIRECHI, S. & LÓPEZ, M. 2009. Aprendiendo por interacción en entornos estocásticos: análisis de performance para algortimos on y off policy. Córdoba: Universidad Tecnológica Nacional.
5. GUTIÉRREZ, P. A., FERNÁNDEZ, J. C. & HERVÁS, C. 2007. Algoritmos de aprendizaje evolutivo y estadístico para la determinación de mapas de malas hierbas utilizando técnicas de teledetección. *II Congreso Español de Informática. IV Taller de Minería de Datos y Aprendizaje.* Córdoba, España: Universidad de Córdoba.
6. IZQUIERDO, E., AMORÓS, J., GÓMEZ, L., MUÑOZ, J., RODRÍGUEZ, J. Z., CAMPS, G. & CALPE, J. 2008. Semi-supervised scheme to update the citric GIS of the Comunidad Valenciana region. *Revista de Teledetección,* 30**,** 23-32.
7. MÉNDEZ, M. C. 2009. *Modelización estadística con Redes Neuronales. Aplicaciones a la Hidrología, Aerobiología y Modelización de Procesos.* Tesis Doctoral, Universidad de Coruña.
8. PRIORE, P., FUENTE, D. D. L., PINO, R. & PUENTE, J. 2002. Utilización de las Redes Neuronales en la toma de decisiones. Aplicación en un problema de secuenciación. *Anales de mecánica y electricidad,* Noviembre-Diciembre.
9. RODRÍGUEZ, E. B., VARGAS-PÉREZ, E., LEYVA-OVALLE, Á. & TERRAZAS-DOMÍNGUEZ, S. 2002a. Aplicación de redes neuronales artificiales y técnicas sig para la predicción de coberturas forestales. *Revista Chapingo. Serie Ciencias Forestales y del Ambiente,* 8**,** 31-37.
10. RODRÍGUEZ, J. R., RIESCO, F. & ÁLVAREZ, C. J. 2002b. Aplicación de la teledetección a la inventariación de viñedo en la Zona de Denominación de Origen Bierzo. *XIV Congreso Internacional de Ingeniería Gráfica.* Santander, España.
11. RODRÍGUEZ, Y., CABRERA, X., FALCÓN, R. J., HERRERA, Z., CONTRERAS, A. M. & GARCÍA, M. M. 2005. CBR-ANN hybrid model to optimize the sequence of wastewater treatments. Universidad Central de Las Villas.
12. SILVA, C., ALVARADO, S., MONTAÑO, R. & PÉREZ, P. 2003. Modelamiento de la contaminación atmosférica por partículas: Comparación de cuatro procedimientos predictivos en Santiago, Chile. Santiago de Chile: Universidad de Chile.
13. TELLO, J. C. 2006. *RE: Reconocimiento de patrones y el aprendizaje no supervisado.*
14. ZHU, X. 2008. Semi-Supervised Learning Literature Survey. *Computer Sciences TR 1530.* University of Wisconsin – Madison.