Propuesta de proyecto de investigación: *Aplicaciones del Aprendizaje Automático en las Ciencias Naturales. Un enfoque interdisciplinario.*

Departamento de Matemáticas Aplicadas y Sistemas
Departamento de Ciencias Naturales
División de Ciencias Naturales e Ingeniería
Universidad Autónoma Metropolitana Unidad Cuajimalpa

15 de noviembre de 2021

Índice

1. Información del proyecto

1.1. Título

Aplicaciones del Aprendizaje Automático en las Ciencias Naturales. Un enfoque interdisciplinario.

1.2. Líneas de investigación

Las líneas de investigación que considera este proyecto son, principalmente, las relacionadas con el los departamentos de Ciencias Naturales y Matemáticas aplicadas y Sistemas, más específicamente, los que tiene que ver con los cuerpos académicos, Optimización, Sistemas complejos e Interfaz Cerebro-computadora, Estudios Moleculares de Sistemas Biológicos.

1.3. Responsable y participantes

UAM Cuajimalpa:

Dr. Roberto Bernal Jaquez (*Responsable*)

Dr. Gerardo Pérez Hernández

Dr. Antonio López Jaimes

Dr. Diego Antonio González Moreno

M.C Luis Ángel Alarcón Ramos

Instituto Nacional de Salud Pública (INSP)

Dr. Gilberto Sánchez-González (*INSP Epidemiología*)

1.4. Orientación

- Investigación básica
- Investigación aplicada
- Desarrollo de Tecnología

1.5. Fecha de inicio y duración

Inicio: 25 de Noviembre de 2021

Duración: 3 años

2. Propuesta

2.1. Resumen

En este proyecto aplicaremos diversas técnicas del Aprendizaje Automático y la Optimización Multiobjetivo en tres importantes problemas de las Ciencias Naturales que, metodológicamente, tienen denominadores comunes: 1. Determinación, clasificación y predicción de las propiedades fisicoquímicas de moléculas usando Redes Neuronales 2. Clasificación de las moléculas semioquímicas que intervienen en la comunicación química interespecie en base a sus propiedades fisicoquímicas y mediante técnicas como Máquinas de Vectores de Soporte (SVM) y Redes Neuronales 3. Predicción de Series de Tiempo usando Aprendizaje Automático y teoría de gráficas usando Redes Neuronales y Máquinas de Regresión con Vectores de Soporte.

2.2. Antecedentes

En los últimos años se ha generado una gran cantidad de información en diversas áreas del conocimiento, como es el caso de la fisicoquímica, donde se tienen Bases de Datos que contiene la información de miles de millones de moléculas como la GDB-19. Esta tendencia de generar grandes volúmenes de información se repite, prácticamente, todas la ciencias naturales y exactas: diariamente se generan miles de terabytes de información en los colisionadores de partículas, en laboratorios de secuenciación biológica, en simuladores mecánicos de alta precisión, en laboratorios de seguimiento epidemiológico, en laboratorios farmacológicos o en estaciones meteorológicos, por citar sólo algunos pocos ejemplos. Este gran flujo de información sería imposible de analizar sin contar con procesos automatizados y confiables [1].
Con éstas premisas es fácil de entender el papel preponderante que en las últimas dos décadas ha tomado el Aprendizaje Automático [2] que nos permite hacer frente a esta tarea y extraer patrones que determinan las características y la dinámica que subyace en el fondo de procesos que suenan absolutamente disímbolos, basando sus conclusiones e inferencias en un estricto conocimiento matemático.

El Aprendizaje Automático que, podemos clasifica en Aprendizaje Supervisado y No-supervisado, basa sus técnicas en la ciencia matemática y su investigación lo lleva a tener un constante crecimiento y se ha enriquecido, especialmente, con la teoría de gráficas, las redes complejas y a últimas fechas, con la computación cuántica.
La abstracción de estos patrones y de la dinámica que determina el comportamiento de los sistemas complejos que describe, se logra usando el paradigma conectivista. Por ejemplo las redes complejas, necesitan grandes cantidades de información para ser entrenadas, pero una vez que se ha realizado su entrenamiento, son capaces de predecir, clasificar y dar las propiedades del objeto de estudio.
En nuestro caso hemos iniciado el estudio de las propiedades moleculares usando un subconjunto de GDB-19. Y entre miles de moléculas podemos clasificar a aquellas que poseen determinadas propiedades fisicoquímicas de interés, predecir los valores de estas propiedades y con ello iniciar lo que podríamos llamar un diseño molecular inteligente usando redes neuronales [3].
En el caso de la comunicación interespecie, se ha observado que, hay una gran variedad de moléculas semioquímicas con diversas propiedades fisicoquímicas. Un problema abierto es determinar cuales son las moléculas involucradas en esta comunicación interespecie y cual es su peso o importancia entre los cientos o miles de moléculas que juegan un rol en la comunicación de plantas o animales. Es aquí donde, el aprendizaje automático puede jugar un papel fundamental puesto que puede clasificar las moléculas de acuerdo a sus propiedades fisicoquímicas que determinan la comunicación y nos dará información de la importancia o peso que determinadas propiedades podría tener.
El aprendizaje automático, posibilita también la predicción de Series de Tiempo, es decir, la determinación de los valores de una o varias variables que, en este caso, varían de una manera no determinista en el tiempo. Este problema puede relacionarse con los dos problemas anteriores en diversas situaciones y permite establecer la evolución dinámica de un sistema.
Nosotros hemos iniciado el estudio de las series de tiempo usando dos casos que podríamos llamar arquetípicos: la predicción en la incidencia de casos de Dengue, que toma en cuenta aspectos climáticos y como segundo caso, la evolución y predicción de patrones melódicos. Es decir, trabajaremos con la identificación y clasificación de un estilo (compositor) y la predicción o composición de una pieza músical que siga un estilo determinado. A mediano plazo, esto permitirá fabricar máquinas que compongan música siguiendo un patrón determinado (el de un compositor).

En este contexto, proponemos un proyecto acorde con (I) las líneas de investigación cultivadas en los Departamentos de la División Ciencias Naturales e Ingeniería y de otras Instituciones (en este caso el Instituto Nacional de Salud Pública) (II) Es propuesto como un proyecto interdisciplinario debido, en gran medida, a que los métodos del Aprendizaje Automático han permeado diversas áreas de conocimiento aprovechando la disponibilidad de grandes Bases de Datos. En particular, podemos observar un gran florecimiento de estudios y resultados en las Ciencias Naturales.

2.3. Objetivos

2.3.1. General

Aplicar diversas técnicas del Aprendizaje Automático y la Optimización Multiobjetivo en tres importantes problemas de las Ciencias Naturales que, metodológicamente, tienen denominadores comunes: 1. Clasificación y predicción de las propiedades fisicoquímicas de moléculas 2. Clasificación de las moléculas semioquímicas que intervienen en la comunicación química inter-especie en base a sus propiedades fisicoquímicas y 3. Predicción de series de tiempo usando Aprendizaje Automático y teoría de gráficas.

2.3.2. Particulares

Ob1: Clasificación y predicción de las propiedades fisicoquímicas de moléculas a partir de un subconjunto de datos de la base GDB-17 usando técnicas de Aprendizaje Automático supervisado (redes neuronales) así como heurísticas de optimización.

Ob2: Clasificación de las moléculas semioquímicas que intervienen en la comunicación química interespecie, en particular, pero no unicamente, en el caso de abejas-ácaros, en base a las propiedades fisicoquímicas de moléculas y usando métodos de aprendizaje no-supervisados.

Ob3: Predicción de series de tiempo usando Aprendizaje Automático y teoría de gráficas tomando como casos de validación a) la determinación de los factores de riesgo e incidencia de los casos de Dengue en el estado de Morelos (con datos que cubren la incidencia en las localidades del estado de Morelos por 10 años, proporcionados por el INSP) y b) la secuenciación de composiciones musicales.

Ob4: Ensayo de diversos modelos y técnicas de aprendizaje automático en problemas de clasificación y predicción así como de técnicas heurísticas de optimización.

Ob5: Estudio de modelos de Aprendizaje Automático que que recurren a gráficas para solución de problemas (Geometric deep learning).

Cabe aclarar que ya se está trabajando en estos objetivos desde hace un año y se tienen resultados preliminares en cada uno de los objetivos particulares.

2.4. Descripción

2.4.1. Hipótesis

Mediante el uso de diversos modelos del Aprendizaje Automático y Bases de datos que poseen un gran número de datos, así como, con la utilización de la teoría de grafos y técnicas heurísticas de optimización, es posible resolver

problemas de clasificación y predicción de propiedades fisicoquímicas, biológicas así como de series de tiempo dando con ello una respuesta algunos problemas e interrogantes de las Ciencias Naturales.

2.4.2. Metodología

Aunque atacaremos problemas que pertenecen a áreas de investigación diversas, la metodología y los modelos que usaremos serán, básicamente, los mismos. Los problemas que atacaremos tratan, entre otras cosas, de predecir ciertas propiedades, por ejemplo fisicoquímicas, y relacionarlas con la estructura molecular o bien, con problemas de clasificación, por ejemplo de moléculas semioquímicas (y sus propiedades fisicoquímicas) que intervienen en la comunicación interespecie.
Daremos un recuento de como serán usados los modelos del aprendizaje automático en cada uno de los problemas planteados:

2.4.3. Clasificación y Predicción de Propiedades Fisicoquímicas de Moléculas

Una parte de la investigación en la que ya tenemos avances es aquella en la que hemos construido Redes Neuronales de retropropagación utilizando como entrada datos de la gigantesca Base de Datos GDB-19 (que contiene la información fisicoquímica de miles de millones de moléculas) o subconjuntos de ella, como la QM-9, y con las cuales hemos sido capaces de predecir las propiedades fisicoquímicas de miles de moléculas, teniendo un entrenamiento previo de la red neuronal con un número reducido de moléculas de la base de datos. Esto es un hito en nuestra investigación y estamos profundizando en usando heurísticas de optimización que permiten un entrenamiento óptimo de la red y la creación de modelos que tienen los parámetros optimizados para hacer las predicciones. Hasta el momento hemos usado como descriptores moleculares a las matrices de coulomb, lo momentos de inercia y algunas de las propiedades fisicoquímicas de las moléculas con las que se realiza el entrenamiento de la red [4].
Lograr estos objetivos un paso fundamental en el diseño de nuevos materiales usando estas nuevas tecnologías.

2.4.4. Factores Fisicoquímicos que influyen sobre la disponibilidad y distribución ambiental de Semioquímicos

En la comunicación interespecie, se ha observado que, hay una gran variedad de moléculas semioquímicas con diversas propiedades fisicoquímicas. Un problema abierto es determinar cuales son las moléculas involucradas en esta comunicación interespecie y cual es su peso o importancia entre los cientos o miles de moléculas que juegan un rol en la comunicación de plantas o animales. Es aquí donde, el aprendizaje automático puede jugar un papel fundamental puesto que puede clasificar las moléculas de acuerdo a sus propiedades fisicoquímicas que determinan la comunicación y nos dará información de la importancia o peso que determinadas propiedades podría tener [5].

Por citar un ejemplo, podemos tomar el trabajo de Lluvia de Carolina Sánchez Pérez y Gerardo Pérez, donde se ha estudiado como el ácaro *Varroa destructor parasita a las larvas de Apis mellifera siguiendo la percepción de los semioquímicos implicados en la comunicación de las abejas; por tanto, la disponibilidad semioquímica y las vías de distribución tienen lugar dentro de diferentes entornos fisicoquímicos. Hemo estudiado la estructura de 172 moléculas con actividad semioquímica sobre Varroa destructor, donde se cuenta con descriptores fisicoquímicos representativos de la partición termodinámica entre diferentes entornos fisicoquímicos: presión de vapor (V), coeficiente de Henry (H), constante de solubilidad en agua (W), coeficiente de partición octanol-agua (O) y coeficiente de partición de carbono orgánico (C); VHWOC.* [1]

En esta parte del proyecto se pretende Aplicar el Aprendizaje de Máquina no supervisado para establecer la tendencia en disponibilidad y distribución de los semioquímicos. Para ello usaremos técnicas como k-means, k-vecinos cercanos y otras técnicas como xgboost. Además, usaremos posteriormente, una red neuronal para tratar de lograr la misma clasificación.

Esta metodología, enriquecida con la gráfica compleja de los actores involucrados en la comunicación interespecies, sentará las bases para lograr avances en el estudio sistemático de la comunicación de plantas y animales del cuál se pueden derivar innumerables aplicaciones como la creación de inhibidores de plagas sin recurrir a pesticidas.

2.4.5. Predicción de series de tiempo. El dengue como un caso de estudio.

Para esta parte de la investigación, se cuenta con una Base de Datos de la incidencia del Dengue en varios estados de la República [6]. En particular, tomaremos los datos del estado de Morelos de donde también tenemos una Base de Datos de las variables climatológicas por 10 años. Preliminarmente, hemos construido una Red Neuronal del tipo LSTM (long short term memory) y un modelo de Regresión de Vectores de Soporte (SVR) que hemos alimentado con nuestros datos. De manera preliminar, hemos obtenido predicciones de gran precisión sobre la incidencia de enfermos de Dengue. Estos modelos involucran también a las variables climatológicas que tuvieron que ser estudiadas para deducir cuales de ellas eran importantes en la incidencia de contagiados [7].

Para mejorar la predicción de secuencias temporales tenemos planeado recurrir a la Teoría de grafos [8] y la auto-codificación de las secuencias temporales para extraer los patrones de la serie y mejorar la predicción o en el caso de secuencias musicales para componer nuevas melodías siguiendo el patrón musical codificado. En esta parte, es importante mencionar que gracias al trabajo del Dr. Diego González Moreno podemos tener encontrar la correspondencia entre la melodía en MIDI que se tiene y un grafo complejo cuyas medidas de centralidad (y otras) ayudan a caracterizarlo.

[1] Tesis de Carolina Sánchez Pérez. Doctorado en Ciencias Agropecuarias UAM-X

Predecir series de tiempo y los factores que determinan la propagación epidemiológica permitirá tomas decisiones que ayuden a mitigar este flagelo de la sociedad mexicana.

2.5. Formación de recursos humanos

En este proyecto se tiene planeada la participación de:

- Al menos un estudiante de posgrado del PCNI a nivel maestría.
- Al menos cinco estudiantes de licenciatura (como Proyecto Terminal)
- Al menos tres estudiantes que realicen servicio social.

2.6. Impacto esperado

El Aprendizaje de Máquina y las grandes Bases de Datos están tomando un auge en todas las esferas de la ciencia y de la vida moderna. Es por ello, este proyecto es de gran impacto pues abre nuevas vías en la solución de problemas científicos de las ciencias naturales.

2.7. Recursos necesarios

2.7.1. Financiamiento

Actualmente **no se cuenta con financiamiento externo**. Se ha logrado avances y resultados preliminares reutilizado equipo de proyectos financiados anteriormente (DMAS). El proceso de investigación avanzará aceleradamente si se cuenta con equipo con procesadores gráficos (GPUs), unidades de almacenamiento de datos y algunas portátiles de alto rendimiento (2 o 3, al menos). Se tiene planeado buscar financiamiento.

2.7.2. Equipo e infraestructura

En este proyecto se tiene contemplado el uso de:

- Equipo de cómputo: Se cuenta con un cluster con 12 nodos para un total de 96 núcleos de procesamiento y se tiene un servidor para cómputo paralelo de 40 núcleos.
- Equipo de Cómputo con procesadores gráficos o GPUs, con los que no se cuenta para realizar esta tareas de investigación (actualmente se tienen algunas GPUs en el DMAS pero dedicados a la docencia).
- Unidad de almacenamiento de 6 Terabytes para almacenar Bases de Datos. No se cuenta con ella.
- De 3 a 5 computadoras personales de alto desempeño.
- Dos puntos de acceso para redes inalámbricas.

- Diversos programas de Software, en su mayoría de Software Libre o de uso gratuito: TensorFlow, Pytorch, Scikit-learn, Julia, JFlux.

3. Calendario de Actividades

A continuación se presenta una planeación del trabajo asociado a los objetivos 1-5 establecidos en la Sección 2.3.2.

Mes	**21-O**	**22-I**	**22-P**	**22-O**	**23-I**	**23-P**	**23-O**	**24-I**	**24-P**
Ob1	*	*	*	*	*	*	*		
Ob2	*	*	*	*	*	*	*		
Ob3			*	*	*	*	*	*	*
Ob4				*	*	*	*	*	*
Ob5					*	*	*	*	*

3.1. Información para el seguimiento

3.1.1. Calendarización de productos esperados

Primer año: trimestres 21-O, 22-I, 22-P

1. Dos Proyectos Terminales.
2. Un estudiante graduado del posgrado de PCNI (maestría).
3. Al menos un artículo indexado. El artículo será sobre la determinación de propiedades fisicoquímicas en moléculas usando GDB-17.
4. Un estudiante de Servicio Social.

Segundo año: trimestres 22-O, 23-I, 23-P

1. Dos Proyectos Terminales.
2. Al menos un artículo indexado. El artículo será sobre la predicción (de series de tiempo) la incidencia del dengue en el estado de Morelos usando técnicas del Aprendizaje Automático. Este artículo será publicado con investigadores del INSP.
3. Un libro (o por lo menos apuntes) de un Curso de Aprendizaje Automático. Ya está en preparación.
4. Un estudiante de Servicio Social.

Tercer año: trimestres 23-O, 24-I, 24-P

1. Un Proyecto Terminal.

2. Dos artículos indexados. Uno versará sobre el Uso de gráficas en la secuenciación y predicción de series de tiempo musicales. El segundo artículo será sobre la Clasificación de semioquímicos involucrados en la comunicación interespecie.

3. Un estudiante de Servicio Social.

3.1.2. Resultados esperados

Aquí haremos un breve sumario de los resultados esperados:

1. **Productividad científica:** En este proyecto se tiene como meta publicar 5 artículos indexados de temática interdisciplinar.

2. **Recursos humanos:** El proyecto dará una formación sólida en Aprendizaje Automático y sus aplicaciones en diversas disciplinas a, al menos, 5 estudiantes de Licenciatura que realicen su Proyecto Terminal y 3 estudiantes de servicio social. Además, formará y graduará a, al menos, un estudiante del posgrado del PCNI.

3. **Impacto:** Este proyecto es interdisciplinario. Como consecuencia será un foro de discusión y análisis para investigadores de diversos departamentos de la UAM y diversas instituciones. Esto es una de los puntos que orientan la actividad de investigación de la UAM Cuajimalpa y constituye una de sus metas.

4. **Impacto:** El rápido avance de la recopilación de Datos y del Aprendizaje Automático han cambiado la manera de trabajar en muchas actividades científicas y tecnológica aparentemente disímbolas: desarrollo de nuevos materiales, descubrimiento de nuevos fármacos, desarrollo de inhibidores de plaga (en lugar de pesticidas) y la predicción de incidencia de enfermedades como el dengue o de secuencias musicales.
   Es importante y estratégico que nuestro país y nuestra universidad, en particular, den un impulso decidido a este singular acercamiento a la solución de problemas. [**?**]

Referencias

[1] J. S. Smith, O. Isayev, and A. E. Roitberg. Ani-1: an extensible neural network potential with dft accuracy at force field computational cost. *Chem. Sci.*, 8:3192–3203, 2017.

[2] F. Berzal. *Redes Neuronales y Deep Learning*. Universidad de Granada, 2018.

[3] Julian Keupp and Rochus Schmid. Topoff: Mof structure prediction using specifically optimized blueprints. *Faraday Discuss.*, 211:79–101, 2018.

[4] Jonathan E. Moussa. Comment on "fast and accurate modeling of molecular atomization energies with machine learning". *Phys. Rev. Lett.*, 109:059801, Aug 2012.

[5] Lluvia de Carolina Sánchez Pérez, Laura Guadalupe Espinosa Montaño, Elizabeth Del Moral-Ramírez, María del Carmen Ramírez-Médeles, Gabriel Gutiérrez-Magdaleno, and Gerardo Pérez-Hernández. Influence of physicochemical factors on environmental availability and distribution of semiochemicals that affect varroa destructor and phylogenetically close organisms: classification by vhwoc pca-clustering. *Heliyon*, 5(8), 2021/11/15 2019.

[6] World Health Organization. Regional Office for South-East Asia. *Comprehensive Guideline for Prevention and Control of Dengue and Dengue Haemorrhagic Fever. Revised and expanded edition.* WHO Regional Office for South-East Asia, 2011.

[7] Gilberto Sánchez-González, Renaud Condé, Raúl Noguez Moreno, and P. C. López Vázquez. Prediction of dengue outbreaks in mexico based on entomological, meteorological and demographic data. *PLOS ONE*, 13:1–14, 08 2018.

[8] Si Zhang, Hanghang Tong, Jiejun Xu, and Ross Maciejewski. Graph convolutional networks: a comprehensive review. *Computational Social Networks*, 6(1):11, 2019.