



Instituto Tecnológico y de Estudios Superiores de
Monterrey

Revisión de artículo: Parte de examen final

Juan Pablo Echeagaray González

A00830646

Diseño de agentes inteligentes

TC2032.101

Juan Emmanuel Ledesma Martínez

18 de marzo del 2022

Resumen

Para esta actividad he escogido el artículo *Distilling Free-Form Natural Laws from Experimental Data* (Schmidt & Lipson, 2009), la idea de usar algoritmos de búsqueda para encontrar leyes físicas que gobiernen nuestra realidad me pareció fascinante.

Una de las tareas más emocionantes y complicadas con la que se enfrenta el ser humano es la comprensión del mundo que lo rodea, el encontrar modelos matemáticos que describan con alta precisión nuestra realidad no es una tarea sencilla; hasta hace un tiempo, esta tarea estaba confinada a científicos o expertos en el área, pero con el surgimiento de algoritmos de búsqueda, y ordenadores con un poder de cómputo enorme, la realización de esta tarea por medio de un computador es de pronto una idea atractiva.

Hay diferentes métodos para aproximar modelos matemáticos que describan la realidad, entre algunos de ellos están los que usan modelos predefinidos, en los que algoritmos de búsqueda deben de encontrar los parámetros que minimicen el error (como una regresión lineal), otros intentan usar algoritmos de búsqueda voraz en espacios monomiales, pero la verdadera meta que queremos alcanzar es la búsqueda de expresiones analíticas libres (sin restricciones) que modelen la realidad.

La regresión simbólica es el nombre de esta tarea. Este proceso se basa en algoritmos de búsqueda evolutivos que reciben como un estado inicial un conjunto de funciones predefinidas por el investigador o desarrollador. Si bien se piensa que la regresión simbólica no es catalogada como un problema con restricciones, esta las tiene; por ejemplo, el modelo no podrá derivar expresiones que no puedan ser generadas a partir del conjunto inicial proporcionado, y tampoco queremos que el modelo agregue cantidades no relevantes a la situación problema. Estas *pseudo* restricciones no entrarían en la clasificación definida por la academia, pero son cosas que el algoritmo debe de considerar.

Como menciono anteriormente, uno puede sesgar los resultados del algoritmo mediante las entradas que uno le proporcione, agregando al tipo de funciones que el agente explorará, están también las variables que podrá usar. Si uno le da solamente posiciones, el algoritmo solo podrá generar ecuaciones en función de estas; si uno proporciona velocidades, el algoritmo estará sesgado a ecuaciones que se enfoquen en energía, si uno le da aceleraciones, se centrará en fuerzas.

Los casos de prueba presentados en el artículo se enfocaron en péndulos, simples y dobles. Al comienzo de la búsqueda, el algoritmo prueba con diferentes combinaciones generadas de su conjunto de variables y funciones iniciales, realiza métricas con las derivadas parciales (aproximadas de los datos capturados) y escoge algunas de las mejores generadas para que estas se *reproduzcan* y pasen sus genes a la siguiente generación, este proceso se repite hasta que las ecuaciones alcancen un nivel de precisión aceptable.

La salida del algoritmo consistió en un juego de las 10 mejores ecuaciones encontradas, al final es tarea del investigador determinar cuál de estas es la mejor. Los investigadores encontraron que dependiendo el número y tipo de variables, el algoritmo arrojaba diferentes tipos de ecuaciones; cuando se le proporcionaron velocidades y posiciones encontró aproximaciones para Leyes de la Conservación de la Energía, cuando se le dieron aceleraciones, formuló la ecuación diferencial que modela la segunda Ley de Newton.

Algunos de los problemas de utilizar estos algoritmos son los siguientes:

- Complejidad computacional: Se requirieron ordenadores de alto poder de cómputo y muchas horas para que el algoritmo convergiera. Este efecto puede reducirse al proporcionarle al algoritmo mejores juegos de ecuaciones, además, estas tareas pueden realizarse en paralelo

- Sesgo: Las respuestas del algoritmo estarán fuertemente sesgadas por el juego de ecuaciones proporcionadas, ¿Cómo se puede determinar cuál es el mejor juego de ecuaciones que no esté sesgado por nuestras percepciones?

Al momento de leer este artículo encontré varias similitudes con los temas vistos en Diseño de Agentes Inteligentes, más algunos conceptos que veo en mi carrera en Ciencias de Datos.

La principal relación que encontramos fue con los algoritmos de búsqueda genéticos, la regresión simbólica implementada por estos investigadores no es más que una excelente aplicación de estos modelos. A la hora de hacer las tareas comprobamos que la velocidad con la que convergían nuestros algoritmos dependía fuertemente de la calidad de las condiciones iniciales que le diéramos. Pero ahora, considerando un caso de la vida real, ¿Cómo podemos determinar una buena condición inicial sin que introduzcamos sesgos en el algoritmo? Es difícil determinar estos estados cuando los algoritmos quieren aplicarse en problemas de la vida real.

Referencias

- [SL09] Schmidt, M., & Lipson, H. (2009). Distilling free-form natural laws from experimental data. *science*, 324(5923), 81-85.