De Luna Ocampo Yanina

Analítica y Visualización de Datos

5CDM1

En las últimas décadas se han visto avances muy importantes dentro de las técnicas de Machine Learning y Deep Learning, sabemos que actualmente día con día se generan millones de cantidades nuevas de información, por lo que éstas son capaces de analizar y aprender de ellas para aplicarlos a ejemplos del mundo real en formatos dispares.

Otra de ellas es la Minería de Datos, que se ha popularizado en las últimas décadas, ésta es la etapa central del proceso de descubrimiento de conocimiento que tiene como objetivo extraer información interesante y potencialmente útil de los datos. Dentro de esta, hay muchas técnicas que pueden ser agrupadas en los siguientes campos: Inteligencia Artificial, Machine Learning, Redes Neuronales y Aprendizaje Profundo.

Relacionado a la minería de datos, tenemos el famoso método abreviado por sus siglas CRIPS-DM, que significa en inglés: Cross-Insdustry Standard Process for Data Mining, que es utilizado para orientar sus trabajos de minería. El ciclo vital de este contiene seis fases con flechas que indican las dependencias más importantes y frecuentes entre fases. La secuencia de las fases no es estricta. De hecho, la mayoría de los proyectos avanzan y retroceden entre fases si es necesario. Es flexible y se pueden personalizar fácilmente.

Otra de las áreas que en la última década ha crecido y se ha puesto en investigación, son las redes neuronales o NN artificiales que son un subconjunto de las técnicas, la aplicación de una de estas técnicas a un conjunto dado de datos puede llevar una cantidad de tiempo considerable, dependiendo de las capacidades informáticas y de almacenamiento que estén disponibles para los científicos. Ofrecen un paralelismo masivo para extender los algoritmos a datos a gran escala por una fracción del costo de un clúster de CPU tradicional de alto rendimiento, lo que permite la escalabilidad sobre conjuntos de datos no computables a través de enfoques paralelos tradicionales. Las técnicas DM se basan en los datos que se analizarán para obtener información de estos datos que proporcionen información relevante para el problema que se está analizando. El cambio en la generación y recopilación de datos también ha llevado a cambios en el procesamiento de datos.

La naturaleza de los datos a gran escala requieres nuevos enfoques y nuevas herramientas que puedas acomodarlos con diferentes estructuras de datos, diferentes escalas espaciales y temporales. El aumento de un gran volumen de información, para ser procesada por la minería de datos y los algoritmos de ML exige nuevas soluciones informáticas paralelas y distribuidas transformadoras capaces de escalar la computación de manera efectiva y eficiente. Las redes totalmente conectadas son la arquitectura más común, ya que se pueden usar para modelas una amplia variedad de problemas que usan datos tabulares.

Siguiendo con el tema de computación acelerada, tenemos que las GPU proporcionan un paralelismo masivo para problemas de DM a gran escala, lo que permite escalar algoritmos verticalmente a datos de volúmenes que no son computables mediante enfoques tradicionales. Los marcos MapReduce con computación GPU pueden superar muchas de las limitaciones de rendimiento y es un desafío abierto para futuras investigaciones. Las soluciones de GPU distribuidas y de GPU múltiples se utilizan para combinar recursos de hardware para escalar a datos más grandes o modelos más grandes.

El paralelismo de datos y el paralelismo de modelos son formas diferentes de distribuir un algoritmo. El paralelismo de datos implica el uso de diferentes nodos para ejecutar la misma porción de código en diferentes lotes de datos. Las ventajas de combinar el paralelismo de datos y el paralelismo de modelos de CNN se pueden encontrar en Krizhevsky. Los fabricantes a menudo ofrecen la posibilidad de mejorar la configuración del hardware con aceleradores de muchos núcleos para mejorar el rendimiento de la máquina/clúster, así como bibliotecas aceleradas, que proporcionan primitivos, algoritmos y funciones altamente optimizados para acceder a la potencia paralela masiva.

Para poder implementar todo esto, tenemos diferentes apoyos de diferentes herramientas y bibliotecas. A continuación, mencionaré algunas y su función:

1. NVIDIA CUDA Deep Neural Network: permite a los usuarios de DL concentrarse en capacitar NN y desarrollar aplicaciones de software en lugar de dedicar tiempo a ajustar el rendimiento de GPU a bajo nivel.
2. OpenMP: es una interfaz de programación de aplicaciones que admite programación de multiprocesamiento de memoria compartida multiplataforma. Su objetivo es facilitar el complicado proceso de análisis de datos y proponer entornos integrados además de los lenguajes de programación estándar.
3. RapidMiner: es una plataforma de software de ciencia de datos de propósito general para la preparación de datos, ML, DL, minería de texto y análisis predictivo.
4. Weka: es un marco popular de código abierto escrito en Java, su propósito está involucrado a un amplio conjunto de algoritmos con esquemas de aprendizaje, modelos y algoritmos.
5. Scikit – Learn: es ampliamente conocido como una popular herramienta Python de código abierto que proporciona funciones para realizar clasificación, regresión, agrupación, reducción de dimensionalidad, selección de modelos y preprocesamiento.

Entre muchas más, que se volvería tardado escribirlas, pero que sin embargo no dejar de ser relevantes en estos procesos, investigaciones y usos.

Aprender y tener en cuenta esto, siempre es importante debido a que, en la actualidad, para los datos es lo que está creciendo y funcionando conforme se va investigando y analizando. Nosotros como analistas de datos, debemos tener conocimiento de esto, debido a que seguro lo utilizaremos en un futuro próximo.