Seminarios del Doctorado en Ciencias Exactas e Ingeniería 2024

Título de Tesis: Inteligencia Artificial y Análisis de Grandes Datos aplicados al estudio de Rayos Cósmicos y Meteorología del Espacio

Tesista: Ms. Ing. Ticiano Jorge Torres Peralta

Directora: Dra. Maria Graciela Molina

Codirector: Dr. Hernán Asorey

Resumen

Los rayos cósmicos son partículas (llamadas partículas primarias) provenientes del espacio exterior que alcanzan la atmósfera terrestre. Estas partículas primarias interaccionan con la atmósfera terrestre produciendo las denominadas cascadas de partículas secundarias denominadas EAS (Extensive Air Showers en sus siglas en ingles). Uno de los métodos más eficientes de detección de estas partículas secundarias a nivel del suelo son los detectores Cherenkov en agua o Water Cherenkov Detectors (WCD). Los WCDs tienen la capacidad de detectar fotones Cherenkov individuales en el rango de 430-570nm, siendo a su vez muy confiables y estables (Sidelnik, I. et al., 2017)). El Latin American Giant Observatory (LAGO)[1] es una red the WCDs colocados por todo Ibero-America, desde México hasta la Antártida. El Tucuman Space Weather Center es uno de los nodos Argentinos de LAGO. En particular en este plan de doctorado se propone estudiar las EAS usando diferentes técnicas de Inteligencia Artificial.

Avances

Los objetivos específicos del plan de trabajo son: a) Estudiar rigurosamente los principales algoritmos de inteligencia artificial (IA), aprendizaje automático en particular, para clasificación; y b) Realizar el estudio de casos relacionados a la meteorología del espacio mediante técnicas de IA a partir de datos y simulaciones dentro de la colaboración. (Por ejemplo, la variación dentro de los diferentes componentes de la cascada en los histogramas de carga).

En este contexto durante el primer año de doctorado se realizaron cursos de postgrado y se avanzo con la investigacion logrando la permitiendo la publicación de un primer trabajo en una revista internacional y actualmente se encuentra en proceso de revisión una segunda publicación.

Respecto a la investigacion cientifica en si, se estudio en profundidad el uso de algoritmos no supervisados (objetivo especifico a) y todos los conceptos asociados para generar un pipeline de procesamiento. Se focalizo en algoritmos no supervisados y específicamente de clustering porque al usar datos reales capturados por un WCD, no esta disponible el ground truth, osea,

1 https://lagoproject.net

no se sabe que partícula secundaria contribuye en particular al histograma de carga. De esta manera se empezó a tratar parte del objetivos específicos b).

Como se acaba de aludir, ademas del algoritmo no supervisado hay una variedad de conceptos necesarios para hacer un modelado apropiado de aprendizaje de maquina (ML). Aquí, el primer paso era profundizar en métodos de limpieza de los datos que incluía: detección y eliminación de anomalías, generación de features, y selección y reducción de features. Esto llevo a un proceso que eliminaba al rededor del 70% de los datos pero aumentaba considerablemente la calidad de set de datos. El restante 39 millones de puntos de datos en el set eran suficientes para analizar. Con esto como primeros pasos del pipeline, se podía continuar a hacer el modelaje.

Específicamente, al explorar una variedad de algoritmos de clustering de diferentes tipos (métodos basados en partición, métodos jerárquicos, métodos basado en densidad, métodos basados en grilla, y métodos basados en distribución), se encontró que métodos de basados en densidad eran los mas apropiados para el tipo de problema. Finalmente se utilizo Ordering Points to Identify the Clustering Structure (OPTICS) (Ankerst, M et al. 1999), un algoritmo jerárquico basado en densidad. Figura 1 encapsula el funcionamiento del mismo.

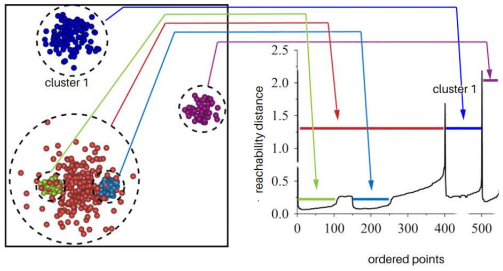


Figura 1: Imagen conceptual mostrando el funcionamiento de OPTICS. Utilizando el Reachability plot producido por el algoritmo, a la derecha, se puede designar thresholds de corte para definir membresía a los clusters, demostrado con las flechas de color.

Resultados iniciales mostraron estructuras muy interesantes en los clusters producidos por OPTICS (Torres Peralta et al., 2023). El algoritmo encuentra claras estructuras entre las features utilizadas, Figura 2 a la izquierda, de las cuales se construyen los clusters en el histograma de carga, Figura 2 a la derecha. El cluster numero 4, por ejemplo, se encuentra en una zona conocida y llamada la joroba de muon (Etchegoyen, A et al. 2005). A pesar de estos resultados preliminares con mucho potencial, utilizando solo los datos reales, no había una facilidad para validarlos.

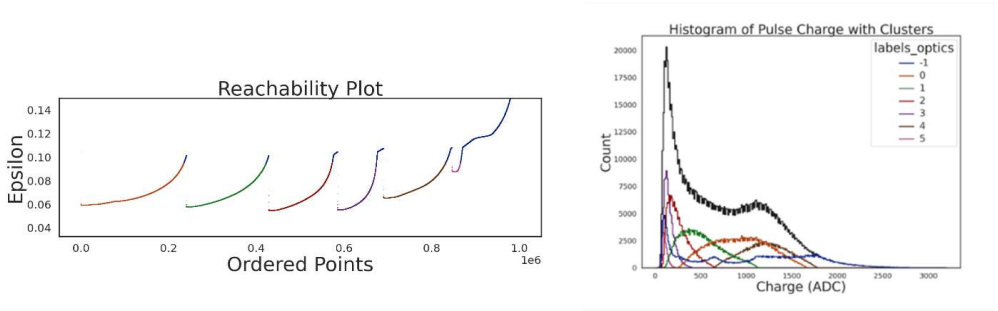


Figura 2: Reachability Plot a la izquierda e Histograma de Carga a la derecha. Los mismos colores se pueden encontrar en los dos graficos para designar los clusters 0 a 5.

Al no tener una forma de validar los resultados con solo los datos reales, se opto por generar datos sintéticos a través de simulaciones. Esto significaba que la simulación tenia que tener en cuenta las mismas características espaciales y atmosféricas durante la fecha de los datos reales y también las mismas características geométricas del WCD real. Este set de datos generado por el suite de simulación de LAGO (Sarmiento-Cano, C. et al. 2022, Taboada, A. et al. 2022), son el que se utilizó.

Como se busca validar el pipeline que se diseño para los datos reales , se utilizo el mismo para los datos sintéticos. Al aplicarlo, se generaron los resultados que se pueden ver en Figura 3. Lo muy interesante es que se pueden ver muchas de las mismas estructuras encontradas con los datos reales, en especial en el sector de la joroba de muon. El beneficio adicional es que en este caso tenemos el 'ground truth' o solución verdadera. La Figura 4 muestra, que el algoritmo es considerablemente bueno agrupando contribuciones muonicas, en amarillo en clusters 0, 1, y 2. También, con las contribuciones de Neutrones (Hadronicas), en verde en cluster 7, pero sin poder separa esta de las electromagnéticas, azul y rosa.

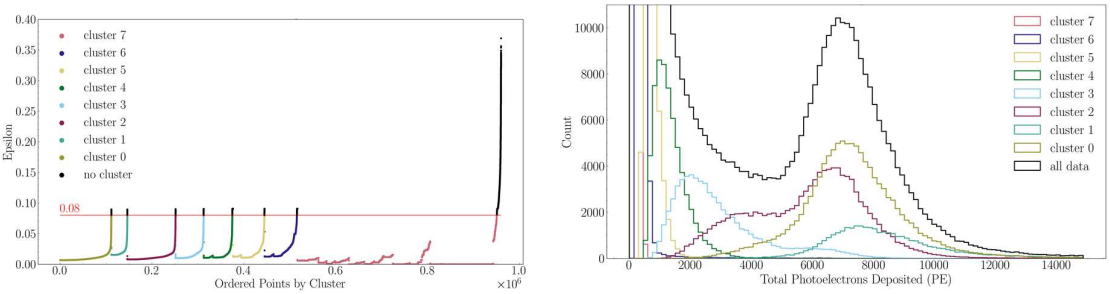
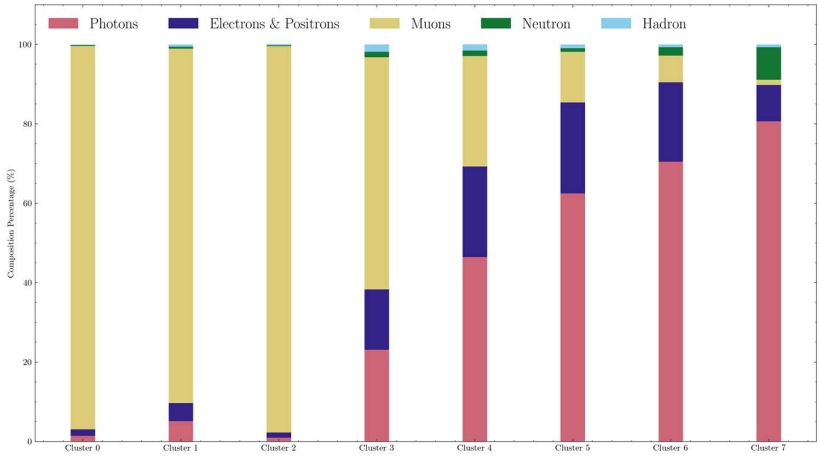


Figura 3: Reachability plot a la izquierda e histograma de carga a la derecha para datos simulados.



.Figura 4: Contribución de partículas secundarias a cada cluster entre el 0 y 7 de izquierda a derecha.

Los resultados con datos simulados validan y dan confianza que el pipeline es muy efectivo para agrupar contribuciones de la clase muonica. Trabajos futuros buscaran perfeccionar el proceso para los otros dos grupos de contribuidores: electromagnéticos y hadrones (en especial neutrones).

Referencias

Ankerst, M.; Breunig, M.M.; Kriegel, H.P.; Sander, J. OPTICS: ordering points to identify the clustering structure. SIGMOD Rec. 1999, 28, 49–60. https://doi.org/10.1145/304181.304187.

Etchegoyen, A.; Bauleo, P.; Bertou, X.; Bonifazi, C.; Filevich, A.; Medina, M.; Melo, D.; Rovero, A.; Supanitsky, A.; Tamashiro, A.; et al. Muon-track studies in a water Cherenkov detector. Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment 2005, 545, 602–612.

Sarmiento-Cano, C.; Suárez-Durán, M.; Calderón-Ardila, R.; Vásquez-Ramírez, A.; Jaimes-Motta, A.; Núñez, L.A.; Dasso, S.; Sidelnik, I.; Asorey, H. The ARTI framework: cosmic rays atmospheric background simulations. The European Physical Journal C 2022, 82, 1019. https://doi.org/10.1140/epjc/s10052-022-10883-z.

Sidelnik, I..; et al. The capability of water Cherenkov detectors arrays of the LAGO project to detect Gamma-Ray Burst and High Energy Astrophysics sources. In Proceedings of the 2022 RICH Conference, these proceedings, Edinburgh, Scotland, 2022.

Sidelnik, I., Asorey, H., Blostein, J. J., & Berisso, M. G. (2017). Neutron detection using a water Cherenkov detector with pure water and a single PMT. Nuclear Instruments and Methods in Physics Research Section A: Spectrometers, Detectors https://doi.org/10.1016/j.nima.2017.02.048

Taboada, A.; Sarmiento-Cano, C.; Sedoski, A.; Asorey, H. Meiga, a Dedicated Framework Used for Muography Applications. Journal for Advanced Instrumentation in Science 2022, 2022. https://doi.org/10.31526/jais.2022.266.

Torres Peralta, T.; Molina, M.; Otiniano, L.; Asorey, H.; Sidelnik, I.; Taboada, A.; Mayo-García, R.; Rubio-Montero, A.; Dasso, S. Particle classification in the LAGO water Cherenkov

detectors using clustering algorithms. Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment 2023, 1055, 168557. https://doi.org/10.1016/j.nima.2023.168557.