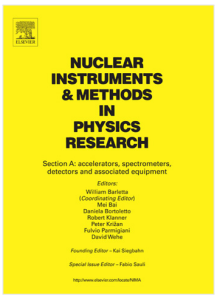
Journal Pre-proof



Particle classification in the LAGO water Cherenkov detectors using clustering algorithms

T. Torres Peralta, M.G. Molina, L. Otiniano, H. Asorey, I. Sidelnik, A. Taboada, R. Mayo-García, A.J. Rubio-Montero, S. Dasso, for the LAGO Collaboration

PII: S0168-9002(23)00547-8
DOI: https://doi.org/10.1016/j.nima.2023.168557
Reference: NIMA 168557

To appear in: *Nuclear Inst. and Methods in Physics Research, A*

Received date : 14 January 2023
Revised date : 11 July 2023
Accepted date : 17 July 2023

Please cite this article as: T.T. Peralta, M.G. Molina, L. Otiniano et al., Particle classification in the LAGO water Cherenkov detectors using clustering algorithms, *Nuclear Inst. and Methods in Physics Research, A* (2023), doi: https://doi.org/10.1016/j.nima.2023.168557.

Particle Classification in the LAGO Water Cherenkov Detectors using Clustering Algorithms

T. Torres Peralta[a,b], M. G. Molina[a,b,c,i], L. Otiniano[d], H. Asorey[g,e], I. Sidelnik[i,f], A. Taboada[g,i], R. Mayo-García[h], A. J. Rubio-Montero[h], S. Dasso[j,k], for the LAGO Collaboration[l]

[a] *Tucumán Space Weather Center (TSWC)*
[b] *Facultad de Ciencias Exactas y Tecnología (FACET-UNT)*
[c] *Instituto Nazionale di Geofisica e Vulcanologia (INGV)*
[d] *Comisión Nacional de Investigación y Desarrollo Aeroespacial (CONIDA)*
[e] *Medical Physics Centro Atómico Bariloche, Comisión Nacional de Energía Atómica (CNEA)*
[f] *Departamento de física de neutrones, Centro Atómico Bariloche, Comisión Nacional de Energía Atómica (CNEA)*
[g] *Instituto de Tecnologías en Detección y Astropartículas (ITeDA)*
[h] *Centro de Investigaciones Energéticas Medioambientales y Tecnológicas (CIEMAT)*
[i] *Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET)*
[j] *Laboratorio Argentino de Meteorología del esPacio (LAMP)*
[k] *Instituto de Astronomía y Física del Espacio (IAFE)*
[l] *The LAGO Collaboration, see the complete list of authors and institutions at https://lagoproject.net/collab.html*

Abstract

The Latin American Giant Observatory (LAGO) is a ground-based observatory studying solar or high-energy astrophysics transient events. LAGO takes advantage of its distributed network of Water Cherenkov Detectors (WCDs) in Latin America as a tool to measure the secondary particle flux reaching the ground. These secondary particles are produced during the interaction between the modulated cosmic rays flux and the atmosphere.

The LAGO WCDs are sensitive to secondary charged particles, high energy photons through pair creation and Compton scattering, and even neutrons thanks to, e.g., the deuteration of protons in the water volume. The pulse shape generated by these particles depends on several factors, such as the detector geometry, the water purity, the sensor response, or the reflectivity and diffusivity of the inner coating. Due to the decentralized nature of LAGO, these properties are different for each node. Additionally, the pulse shape depends on the convolution between the response of the central photomultiplier (PMT) to individual photons and the time distribution of the Cherenkov photons reaching the PMT. Typically, a WCD gives pulses with a sharp rise time ($\sim$ 10 ns) and a longer decay time ($\sim$ 70 ns).

In this work, the WCD data used is acquired using the original LAGO data-acquisition system that digitizes pulses at a sampling rate of 40 MHz and 10 bits resolution on time windows of 400 ns. Here, we apply unsupervised machine learning techniques to find patterns in the WCDs data and subsequently create groups, through clustering, that can be used to provide particle separation. We use data acquired from an individual WCD, showing that density-based clustering algorithms are suitable for automatic particle separation producing good candidate groups. Improved separation would help LAGO to reconstruct *in situ* the properties of primary cosmic rays flux. These results open the possibility to deploy machine learning-based models in our distributed detection network for onboard data analysis as an operative prototype, allowing detectors to be installed at very remote sites.

Keywords: Principal Component Analysis, OPTICS, Machine Learning, Water Cherenkov Detector

1. Introduction

Astroparticles that constantly impinge on the Earth's atmosphere are the reason for the existence of an atmospheric flux of secondary particles composed of three main components: the electromagnetic (γs and $e^{\pm}$), the muonic ($\mu^{\pm}$) and the hadronic (composed of different types of mesons and baryons, including nuclei).

The Latin American Giant Observatory[1] (LAGO) consists of a network of Water Cherenkov Detectors (WCD) located at various sites in Latin America. Some of LAGO's principal objectives include the measurement of events that have extreme energy coming from space with the use of WCDs at ground-level sites [1], and the continuous improvement of our WCD systems [2]. LAGO WCDs use a single large-area photomultiplier tube as the primary sensor. When ultra-relativistic charged particles cross the WCD they cause Cherenkov radiation, which in consequence triggers a detection in the data acquisition system of the detector. Due to its large water volume, neutral particles, such as photons or neutrons, can also be indirectly detected through Compton scattering or pair creation of the former case, or nuclear interactions with the different materials present in the latter case.

This work aims to identify each of the components detected

[1] https://lagoproject.net

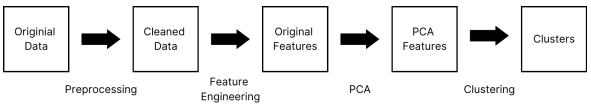


Figure 1: Data processing pipeline. Raw data is passed through a pre-processing stage to remove anomalies; the feature engineering stage extracts information from the cleaned data; the feature selection stage uses PCA to select principal components as new features; finally, the clustering stage uses OPTICS to cluster the data points.

by LAGO WCDs by the application of an unsupervised machine learning (ML) pipeline to find patterns in the data and to create groups through clustering.

The pipeline consists of several stages including an aggressive data pre-processing stage to clean the dataset because of the inherent characteristics of the raw data. Next, there is a feature engineering stage that extracts information from the cleaned dataset. This is followed by a feature selection stage that employs Principal Component Analysis (PCA) to identify relevant features. Finally, a machine learning modelling stage utilizes a density-based clustering algorithm called Ordering Points To Identify Clustering Structure (OPTICS) to cluster the captured particles from the WCD. Figure 1 depicts this pipeline.

2. Raw data pre-processing

Raw data used in this work is provided by the LAGO's "Nahuelito" WCD site at Bariloche, Argentina. Pulses that are captured by the data acquisition system (DAQ) are digitized at a sampling rate of 40 MHz and a 10 bits resolution on time windows of 400 ns. A total of 24 hours of raw data is used starting at 13:00 ART (10:00 UTC) on March 01, 2012.

The data used is the raw data as captured by the DAQ, thus pre-processing is an essential stage in the overall data pipeline. Here, we cleaned the data from electronic noise and anomalies as it is the standard procedure before any analysis. Due to the nature of the electronic system and the WCD characteristics, the original data presents a high number of anomalies, about 60% of the total. Five types of anomalies are defined as can be seen in Table 1. All five types are eliminated, resulting in a clean data set of 39 million pulses (data points). We consider that this is enough data points in relation to the number of features used and the number of clusters found, although there is no general rule for the minimum number of data points required for clustering analysis [3].

3. Feature Engineering and Selection

With regards to feature engineering, the set of proposed features is summarized in Table 2. In general, one must assure to have both enough information and a sufficiently low number of dimensions [4]. For this reason, we applied a standard procedure of normalizing the features before PCA, which takes the original features as input and results in a new set of orthonormal features called principal components [5]. This new set is used in the subsequent clustering stage.

Table 1: Types of pulses that are considered anomalous and are eliminated. These constitute about 60% of the 98 million pulses from the original data set. Each type is defined and its percentage of the total anomalous pulses is shown.

Name	Definition
Saturated Pulse (∼1%)	A pulse with any sample reaching the saturation ADC peak value (1023).
Duplicated Pulse (< 0.1%)	A pulse whose ID and values are exactly the same as another.
Complex Pulse (∼21%)	A pulse that did not have the expected Fast-Rising-Exponential-Decay form.
Negative Pulse (∼1%)	A pulse where the DAQ reported negative numbers.
Short Pulse (∼77%)	A noise associated pulse that triggered the acquisition (third temporal bin) but did not surpass a secondary threshold of 70 ADC on the fourth temporal bin [2].

Table 2: Original features used before the application of PCA.

Name	Description
Charge (Area)	Total charge deposited (the time integration of the pulse)
Peak	Maximum value of the pulse.
Charge Deposit Time	Number of time bins from the triggered bin to the peak bin.
Width	Number of bins over the detection threshold
Delta Time Forward	Time difference between current and next pulse.
Delta Time Backwards	Time difference between current and previous pulse.

4. Method

The ML model is the hierarchical density-based clustering algorithm OPTICS, which is an unsupervised ML technique [6]. Like all clustering algorithms, its objective is to group similar points in the data set, which in our context means pulses belonging to the same type of particle. Specific to density-based algorithms, clusters are chosen according to regions of high density defined by a fixed neighbourhood distance epsilon (ε), and are separated by regions of low density.

What is particular to OPTICS is that it defines the reachability distance, a minimum distance (ε) between a particular point and the closest core group. Each point that connects to a core group is ordered from smallest to biggest reachability for that particular group. This unique strategy creates a reachability plot where each section represents a core group. This captures information about every cluster level present as can be seen in Figure 2.

In the reachability plot, areas of low ε are considered "valleys" and correspond to places where there is a higher density of data points, meanwhile, areas where the value of ε rises more vertically, are considered the edges of the "valley". A visual strategy can be used to choose a maximum ε value as a threshold to decide if a given point is a member of a given cluster, as proposed by [6].

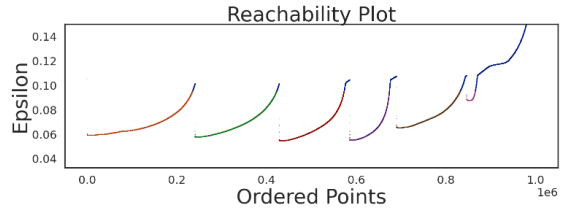


Figure 2: Reachability plot. Each cluster (represented by each colour) is composed of those points of higher density ("valley"). Points in the edges of the valley that are above the defined threshold of 0.095 have less density and do not belong to any group (blue).

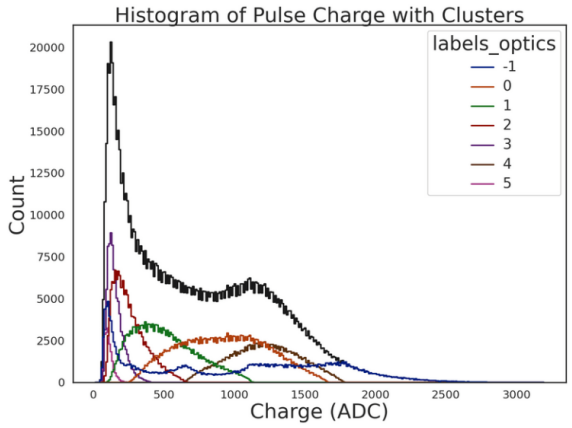


Figure 3: Charge histogram of the dataset with the six detected clusters (0-5) and the points that do not belong to any cluster (-1). Cluster 4 is of particular interest as it is located where VEMs are expected to be found. Clusters 2 and 5 are also of interest as they are located where electromagnetic contributions are expected to be found.

5. Results

With the use of the visual strategy, as mentioned in the previous section, a conservative threshold of 0.095 is used. Nevertheless, in future works, we will perform fine-tuning of several hyperparameters including this threshold. This results in six clusters as seen in Figure 3. Any point above the threshold is regarded as not being part of any cluster (labelled -1 in blue).

Based on the well-known response of WCD for different types of component of extensive air showers, which is characterised by the position and the features observed for each cluster in the charge histogram (Fig. 3), it is possible to infer some possible identification for the obtained clusters. For example, cluster 4 (brown) could correspond to the Vertically Equivalent Muon (VEM) contributions as they are expected to be found in that particular region of the histogram. In the same way, clusters 3 (purple) and 5 (pink) could be associated with electromagnetic contributions as the total deposited energy is directly related to the total energy of each particle, and so, they are expected to be found on the left side of the charge histogram [7].

These results show that is possible to group the contributions of the components in separated clusters. Preliminary validation was done by visually inspecting the clusters. Nevertheless, further systematic validation is needed which is planned in future works by using synthetic data under controlled conditions.

6. Conclusion

The OPTICS clustering algorithm produced promising results. Cluster groups are located where secondary particle contributions are primarily expected to appear (e.g. muonic and electromagnetic). The reachability plot shows clear cluster structures based on the density of the features used.

These preliminary results can be a starting point for future steps in this research including further efforts to improve the pre-processing procedures [2] and study the feature space, the application of hyper-parameter tuning techniques, and additional validation of results with the use of simulated data [7] and actual labelled data that are available for specific cases.

It is worth mentioning that, recently, many works have applied machine learning algorithms to synthetic data (e.g. obtained from simulations), thus we consider that analyzing actual data is an important contribution of this work.

In the long term, this research is expected to be implemented in a semi-real-time manner directly onboard the WCDs of the LAGO Collaboration detection network.

7. Acknowledgments

This work was partly carried out within the 'European Open Science Cloud - Expanding Capacities by building Capabilities' (EOSC-SYNERGY) project, co-funded by the European Commission's Horizon 2020 RI Programme under Grant Agreement nº 857647. We acknowledge the ICTP and OIEA grant NT-17 that partially funded stays to carry out this work. The LAGO Collaboration is very thankful to all the participating institutions and to the Pierre Auger Collaboration for their continuous support.

References

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

[2] L. Otiniano, et al., Measurement of the Muon Lifetime and the Michel Spectrum in the LAGO Water Cherenkov Detectors as a tool to improve energy calibration and to enhance the signal-to-noise ratio., in: 2022 RICH Conference, these proceedings, Edinburgh, Scotland, 2022.

[3] K. Siddiqui, Heuristics for sample size determination in multivariate statistical techniques, World Applied Sciences Journal 27 (2013) 285–287.

[4] E. Debie, K. Shafi, Implications of the curse of dimensionality for supervised learning classifier systems: theoretical and empirical analyses, Pattern Analysis and Applications 22 (2019) 519–536. doi:10.1007/s10044-017-0649-0.

[5] I. T. Jolliffe, J. Cadima, Principal component analysis: a review and recent developments, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 374 (2016) 20150202. doi:10.1098/rsta.2015.0202.

[6] M. Ankerst, et al., OPTICS: ordering points to identify the clustering structure, ACM SIGMOD Record 28 (1999) 49–60. doi:10.1145/304181.304187.

[7] C. Sarmiento-Cano, et al., The ARTI framework: cosmic rays atmospheric background simulations, The European Physical Journal C 82 (2022) 1019. doi:10.1140/epjc/s10052-022-10883-z.