PROCEEDINGS OF SCIENCE

ONLINE ICRC 2021
THE ASTROPARTICLE PHYSICS CONFERENCE
Berlin | Germany
37th International Cosmic Ray Conference
12–23 July 2021

Model independent search for transient multimessenger events with AMON using outlier detection methods

T. Grégoire,[a,*] H. A. Ayala Solares,[a] S. Coutu,[a] D. Cowen,[a] J. J. DeLaunay,[a] D. B. Fox,[a] A. Keivani,[b] F. Krauss,[a] M. Mostafá,[a] K. Murase,[a] E. Neight[a] and C. F. Turley[a] for the AMON group

[a] *Pennsylvania State University, Department of Physics*
State College, USA
[b] *Columbia University, Department of Physics,*
New York, USA
E-mail: tmg5746@psu.edu, hza53@psu.edu, sxc56@psu.edu, jjd330@psu.edu, dbf11@psu.edu, azadeh.keivani@columbia.edu, felicia.krauss@psu.edu, mam1264@psu.edu, kum26@psu.edu, cft114@psu.edu

The Astrophysical Multimessenger Observatory Network (AMON) receives subthreshold data from multiple observatories in order to look for coincidences. Combining more than two datasets at the same time is challenging because of the range of possible signals (time windows, energies, number of events). However, outlier detection methods can circumvent this issue by identifying any signal divergent from the background (e.g. scrambled data).

We propose to use these methods to make a model independent combination of the subthreshold data of neutrino and gamma ray experiments. Using the python outlier detection (PyOD) package, it allows us to test several methods from a simple "k-nearest neighbours" algorithm to a more sophisticated Generative Adversarial Active Learning neural networks which generates data points to better discriminate inliers from outliers.

37th International Cosmic Ray Conference (ICRC 2021)
July 12th – 23rd, 2021
Online – Berlin, Germany

*Presenter

PoS(ICRC2021)934

1. AMON

The last decades have seen the emergence of multimessenger astrophysics. Indeed, the universe is now studied through the observation of photons, cosmic rays, neutrinos as well as gravitational waves and the combined observations of a source from multiple messengers has proven to be enlightening. The coincident detection of gravitational waves and electromagnetic radiations allowed the first detection of the coalescence of a binary neutron star [1]. Multiple messengers bringing different information are very instructive when put together. That is also the case of the first evidence of a high energy neutrino source [2, 3] from the coincident detection of neutrinos and a gamma ray flare.

The Astrophysical Multimessenger Observatory Network (AMON) [4] has been developed at the Pennsylvania State University, with the goal to combine subthreshold data from several astrophysical observatories near realtime. Indeed, AMON has signed Memoranda of Understanding (MoU) with different collaborations in order to receive their data below the discovery threshold in real-time. Currently IceCube and ANTARES send subthreshold track events to AMON. We also receive IceCube high energy events above the detection threshold, the "Gold" and "Bronze" track events [5] as well as the cascades [6]. HAWC sends two real-time datasets to AMON, the "hotspots" and the "bursts" [7]. Fermi-LAT data are also stored in the AMON database. All the data received are stored in the AMON servers in order to do archival analyses.

The subthreshold data are background dominated and cannot be used to identify a signal alone, however if an excess is detected by multiple instruments their combined signal can become significant. The AMON team is looking for such signal by analysing the data in real-time thanks to the AMON infrastructure. AMON sends any statistically significant result publicly to the Gamma-ray Coordinate Network (GCN) so that small field-of-view instruments can point toward the direction of the signal, looking for a counterpart. The data received are also stored in the AMON servers in order to do archival analyses.

By doing so, AMON contributes to the search and study of the most energetic phenomena in the universe, helping its partners to best exploit their data in order to answer fundamental questions of astrophysics, fundamental physics and cosmology.

The analysis presented in this proceeding aims at combining more than two datasets at the same time, which is something AMON was specifically designed to do.

2. Multiple Datasets Outlier Detection

We present here a search for coincident signal in several datasets based on outlier detection methods. Indeed, the search combining several datasets must allow the detection of a large range of signal coincidences, and it would not be feasible to simulate realistically all possible signal combinations and quantify the probability of one combination in respect to an other one. Outlier detection methods permit an agnostic search for signal by learning the background in order to classify any divergent data point as signal. In contrast to simulating signal, simulating background by scrambling the data is a straight forward task.

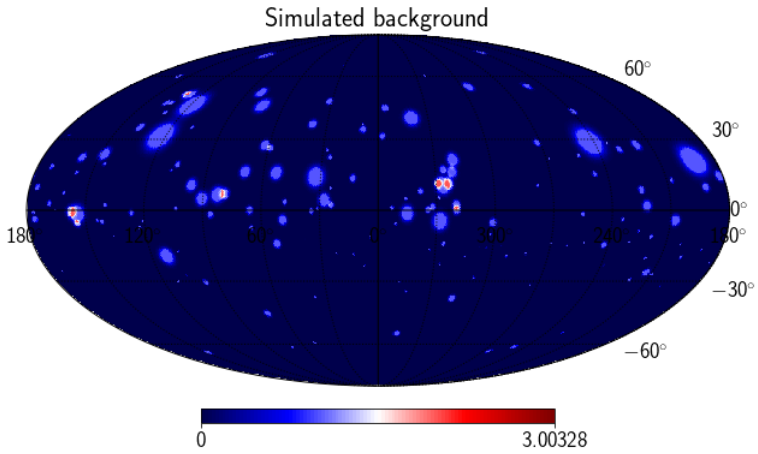

Figure 1: Skymap of the event density of a simulated background for a dataset with a large range of angular error sizes, for illustration purpose.

The method presented here is mostly independent from the datasets used as inputs. It could be applied to data currently available to AMON as well as future new data. We present the method in the form of an archival search, but it could also be used for a real-time stream in the future.

2.1 Input Data

This analysis takes as input a list of events with their corresponding date, position and position uncertainty for each dataset. These data are converted into skymaps of event densities for each time steps of 6h as illustrated in Fig. 1. To avoid that two events close in time fall within different time windows, the skymaps are done twice with a 3h shift in time.

The event density is defined as being unity for pixels within the 68% error region of the event and it decreases following a 2D Gaussian for larger distances, as shown in Fig. 2. The Gaussian is scaled to be unity at the 68% error contour $d_{68\%}$ for continuity, for a 2D Gaussian $d_{68\%} \approx 1.515\sigma$.

$$\text{event density} = \begin{cases} 1, & \text{if } d < d_{68\%} \\ \exp\left(\dfrac{-d^2 + d_{68\%}^2}{2\sigma^2}\right), & \text{otherwise} \end{cases}$$

This event density is chosen as it gives a larger value in the case of a pixel surrounded by a few events (e.g., an event density = 3 for 3 nearby events) than in the case of only one centred event (= 1) while the use of a Gaussian would not always allow one to distinguish a pixel with several nearby events from a pixel containing only one centered event. However, we plan to test different event density definitions.

The skymaps of the event densities are used to build the input data of the outlier detection algorithm. Each data point corresponds to a pixel of a time step and contains n event densities corresponding to the n datasets to combine, as well as the altitude and azimuth of the pixel seen from the so-called "Null Island" corresponding to the 0° N, 0° E in Earth coordinates.

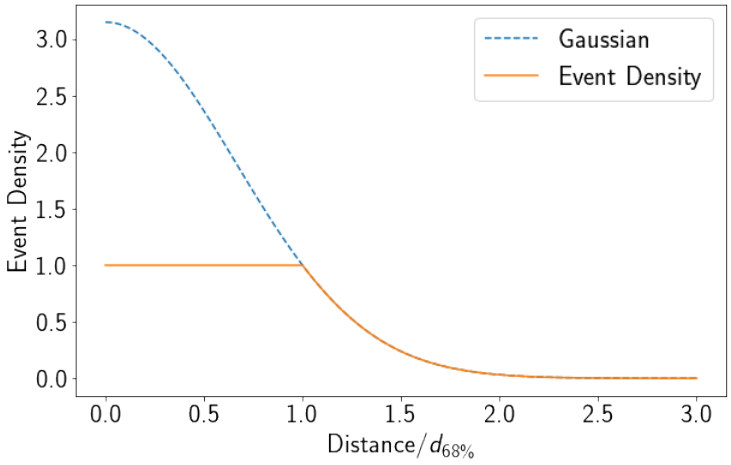

Figure 2: Event density of an event in a pixel as a function of the distance.

2.2 Scrambling

The training of the outlier detection algorithm should be done on background only, therefore a scrambling of the data is done to simulate background. The scrambled data are also used to blind the analysis. The scrambling is typically done by permuting the azimuth and time of the events of a dataset. The equatorial coordinates are then computed from the scrambled local coordinates.

2.3 Outlier Detection Algorithms

Several outlier detection algorithms exist, therefore the PyOD (*Python Outlier Detection*) [8] library was used as it implements many algorithm. We will test several of them and choose the one that fits best to our needs. Here is a short description of some of these algorithms:

- *K-nearest neighbours (KNN)* [9, 10]: One of the simplest algorithm but still effective. The outlier score of a data point is its distance to its k^{th} nearest neighbour.
- *Histogram-based Outlier Score (HBOS)* [11]: Very fast algorithm but less precise as it assumes the independence of the features.
- *Principal Component Analysis (PCA)* [12]: Consists in a decomposition of correlated variables into a lower dimensional space of uncorrelated variables, the so-called "principal components". The outlier score of a data point is the sum of its projected distances on the principal components. However, in our case the input data are made of a few uncorrelated variables as the background of the different datasets are independent, while PCA is best for cases when the input is made of many partially correlated variables.
- *AutoEncoder* [13]: A neural network that learns to encode a set of data into a lower dimension space and decode it back to its initial values. The error between the input and the output should be small only if the input is similar to the training sample, therefore the outlier score is the reconstruction error. The autoencoder gives similar results to the PCA. This is expected as autoencoders are a nonlinear extension of PCA.

- *Multiple Objective Generative Adversarial Active Learning (MO-GAAL)* [14]: One of the most advanced algorithms, using neural networks. It is composed of a discriminator and multiple generators. The generators try to imitate the data as best it can and the discriminator tries to distinguish the data, considered as inliers, from the generated events, considered as outliers.

PyOD also allows to use a combination of multiple algorithms.

2.4 Signal Injection

In order to choose the algorithm that will be the most sensitive, signal events are simulated. However, as stated previously it is not possible to have a representative simulation of all the possible signals, therefore this simulation is used for a proof of concept and to choose between multiple algorithms, but it does not allow us to get the sensitivity of the analysis to any signal.

The signal injection is done by picking a random direction and time and injecting signal in three or more of the datasets at this location accounting for the event's angular uncertainty and the detector's field-of-view.

2.5 Output

The outlier detection algorithm outputs an outlier score for each pixel of the skymap at each time step. However a signal is usually larger in extent than a single pixel and therefore adjacent pixels with a high outlier score are combined into one signal event. The outlier score of the event is the maximum score of its pixels.

3. Status and Perspectives

This analysis is approaching maturity, and we plan to use it on archives of five datasets AMON receives [4]: the ANTARES tracks, IceCube singlets, HAWC hotspots and HAWC bursts as well as Fermi LAT realtime data.

This analysis could be run in realtime in the future in order to send alerts and trigger follow-ups of the most significant outlier events. More datasets could also be added in the future without having to develop a new analysis.

References

[1] B. P. Abbott *et al.* *The Astrophysical Journal* **848** no. 2, (Oct., 2017) L13.

[2] M. G. Aartsen *et al.* *Science* **361** no. 6398, (July, 2018) 147–151.

[3] The IceCube Collaboration, Fermi-LAT, Magic, Agile, Asas-Sn, Hawc, H.e.s.s, Integral, Kanata, Kiso, Kapteyn, Liverpool Telescope, Subaru, Swift/NuSTAR, Veritas, and Vla/17b-403 Teams *Science* **361** no. 6398, (July, 2018) eaat1378.

[4] H. A. Ayala Solares, S. Coutu, D. Cowen, J. J. DeLaunay, D. B. Fox, A. Keivani, M. Mostafá, K. Murase, F. Oikonomou, M. Seglar-Arroyo, G. Tešić, and C. F. Turley *Astroparticle Physics* **114** (Jan., 2020) 68–76.

[5] http://gcn.gsfc.nasa.gov/doc/IceCube_High_Energy_Neutrino_Track_Alerts_v2.pdf.

[6] https://gcn.gsfc.nasa.gov/doc/High_Energy_Neutrino_Cascade_Alerts.pdf.

[7] https://gcn.gsfc.nasa.gov/doc/hawc_grb_alerts.pdf.

[8] Y. Zhao, Z. Nasrullah, and Z. Li *Journal of Machine Learning Research* **20** no. 96, (2019) 1–7.

[9] F. Angiulli and C. Pizzuti, "Fast Outlier Detection in High Dimensional Spaces," in *Principles of Data Mining and Knowledge Discovery*, T. Elomaa, H. Mannila, and H. Toivonen, eds., Lecture Notes in Computer Science, pp. 15–27. Springer, Berlin, Heidelberg, 2002.

[10] S. Ramaswamy, R. Rastogi, and K. Shim *SIGMOD Rec.* **29** no. 2, (May, 2000) 427–438.

[11] M. Goldstein and A. Dengel, "Histogram-based Outlier Score (HBOS): A fast Unsupervised Anomaly Detection Algorithm," pp. 59–63. KI-2012: Poster and Demo Track, Sept., 2012.

[12] M.-L. Shyu, S.-C. Chen, K. Sarinnapakorn, and L. Chang, "A Novel Anomaly Detection Scheme Based on Principal Component Classifier," in *Proceedings of International Conference on Data Mining*. Jan., 2003.

[13] C. C. Aggarwal, *Outlier Analysis*. Springer-Verlag, New York, 2013.

[14] Y. Liu, Z. Li, C. Zhou, Y. Jiang, J. Sun, M. Wang, and X. He *IEEE Transactions on Knowledge and Data Engineering* **PP** (Mar., 2019) 1–1.

PoS(ICRC2021)934