PROCEEDINGS OF SCIENCE

Gamma Hadron separation using single parameter method and multivariate algorithms with LHAASO-WCDA experiment

X.J. Wang*[a], W.Y. Liao[b], Z. Cao[a], M. Zha[a], Z.G Yao[a], H.R. Wu[a] for the LHAASO Collaboration

[a] *Institute of High Energy Physics, Chinese Academy of Sciences, Beijing, 100049, China*
[b] *University of Nankai, Tianjin 300071, China*
E-mail: wangxiaojie@ihep.ac.cn

The Large High Altitude Air Shower Observatory (LHAASO) is built at an altitude of 4410 meters in Daocheng, Sichuan Province, China. As one of its major components, the Water Cherenkov Detector Array (WCDA) will focus on surveying the Northern sky for gamma-ray sources in a wide energy range. Because most of the triggered events are induced from hadronic cosmic-rays, suppressing a large number of background events is very important for LHAASO-WCDA. In this work, several sensitive parameters are chosen for the single parameter gamma/hadron discrimination based on topology characteristics of showers. Different methods of multivariate analysis are also used in this study. Results from simulation data show that multivariate analysis can improve the separation significantly compared to single traditional methods. Some preliminary results using the separation technique on experimental data are introduced.

36th International Cosmic Ray Conference -ICRC2019-
July 24th - August 1st, 2019
Madison, WI, U.S.A.

*Speaker.

PoS(ICRC2019)820

1. Introduction

As a messenger gamma-ray can provide more information about their source than others because of its electric neutrality. The Water Cherenkov Detector Array(WCDA) of Large High Altitude Air Shower Observatory(LHAASO) is a ground-based gamma-ray observatory which will focus on surveying the northern sky for steady and transient sources from 100 GeV to 30 TeV. LHAASO-WCDA is composed of three water ponds with the area of 78,000 m^2, the effective water depth of 4 m, to be constructed at 4410 meters a.s.l in Daocheng, Sichuan Province, China [1]. Every pond is divided into cells with a size of 5m x 5m, partitioned by black plastic curtains to prevent penetration of the lights yielded in neighboring cells. Every cell has two photo-multiplier tubes(PMTs) at the bottom to collect Cherenkov lights generated by charged secondary particles of extensive air shower, see as Fig 1. Since 2017, one of the three ponds has been built, equipped with an 8 inch PMT and 1.5 inch PMT. For the sake of better sensitivity at low energies in the upgrade plan, for the other two ponds.

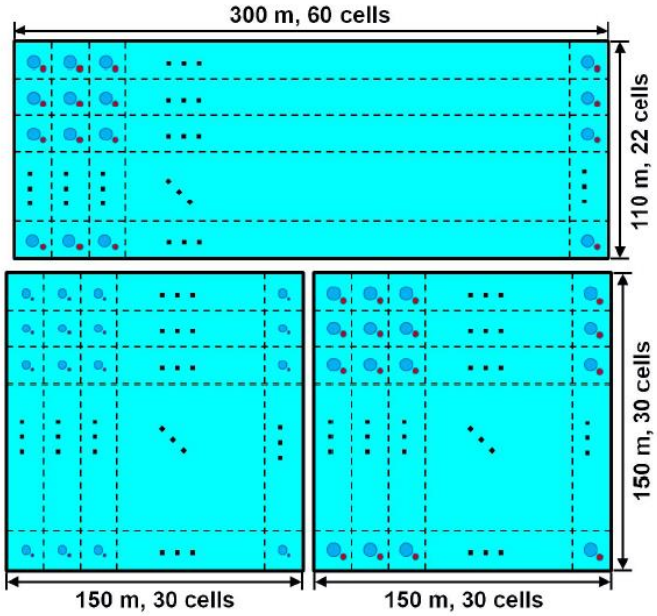


Figure 1: Sketch drawing of WCDA layout

Given that most of the events detected by WCDA are induced by cosmic rays, the study of gamma-ray sources has an urgent need for effective methods to reject the background of cosmic rays. Both of the single parameter methods and multivariate methods are used to achieve a good enough gamma/hadron separation. They are to be introduced in this paper. Since one pool has been built and beginning to take data from April, all the results are based on 900 cells.

2. Simulation data

The air shower events are generated with CORSIKA v75000 [2]. The QGSJET-II model [3]is adopted for high energy hadronic interactions, and the FLUKA is used for low energy ones. In this

generation, 5 different primary cosmic ray nuclei (proton/helium/CNO/MgAlSi/Fe) are used, their fluxes are from the Horandel compilation [4]. The energy range for all types of primary particles is set to 10 GeV- 100 TeV, but it is segmented into several ranges such that every energy range can have some events. Samples in different energy ranges are combined by applying a series of proper weighting factors.

The tracking of secondary particles in the WCDA detector array is simulated with a code based on Geant4 [5] and the PMT model are taken from GenericLAND software library [6]. The quality of pure water in the pool has a great effect on the experiment. According to a study, if the water transparency is not so good(attenuation length is less than 40m), it will lead to absorption of the visible lights [7]. In this work, the attenuation length of water is 8 m. Additional details on the simulation data are available in [8].

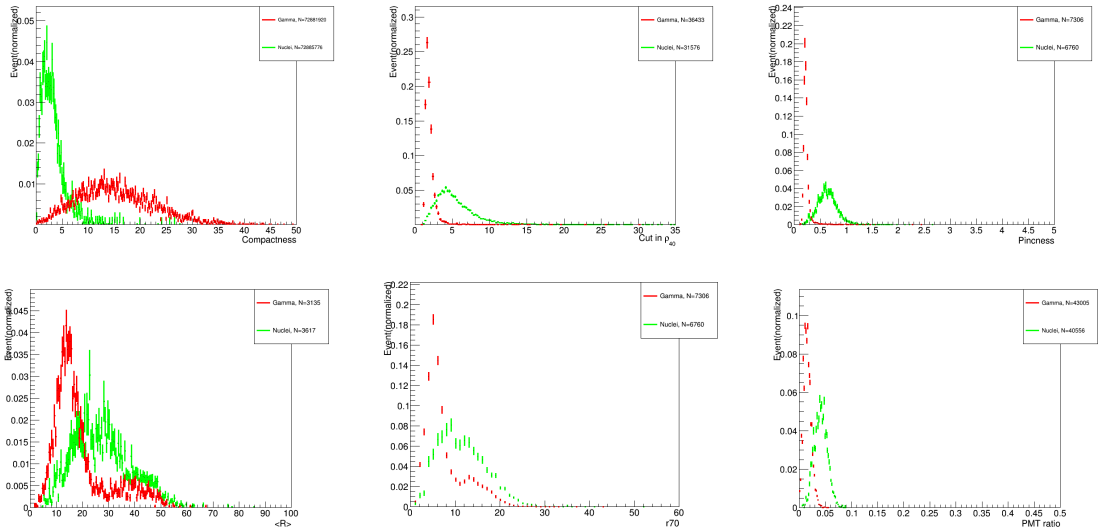


Figure 2: The distributions of six mass sensitive parameters, *Compactness*, ρ_{40}, *Pincness*, $< R >$, $r70$ and $PMTRatio$ for gamma (red line) and nuclei (green) initialed showers.

3. Sensitive parameters

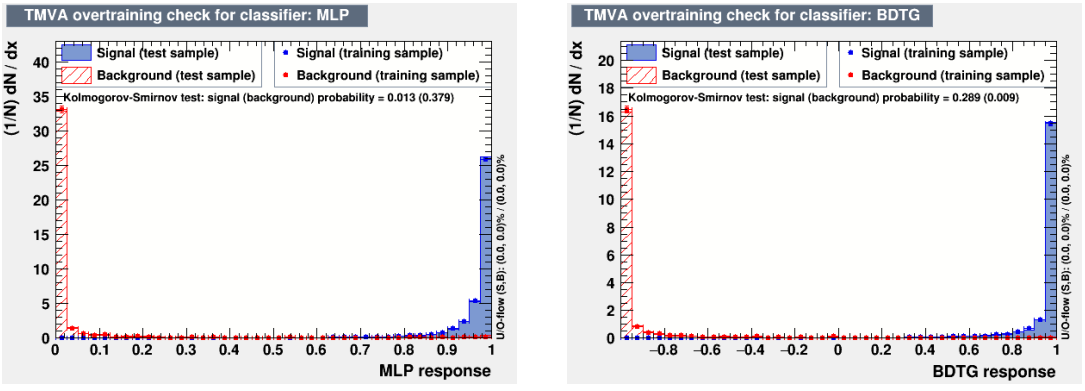
Because of the difference in main interact process during the development of gamma induced air shower and hadronic induced shower, they can be recognized by their topological character. The gamma-like shower often has a compact, smooth lateral shower profile while the hadronic-like showers are more cluttered, irregular, larger signals farther from the shower core will be generally found. According to the information above, several parameters are used to identify cosmic-ray events and their separation power are shown in Fig2.

- The first parameter is called compactness, C , its definition is $C = nFit/CxPE_{45}$. Where nFit is the number of PMTs during the reconstruction process.$CxPE_{45}$ is the largest PE of fired PMTs outside a radius of 45 m from the reconstructed shower core. Usually, if a shower event has a bigger C, it is more likely a gamma event.

- The second parameter is named as ρ_{40}. It is the average density outside 40 m from event core. $\rho_{40} = \sum PE_{40} / \sum PMT_{40}$
- The third parameter is <R>, mean lateral spread radius of particle flow from shower core, where $< R >= \sum(PE_i R_i) / \sum PE_i$. Where PE_i and R_i are the numbers of detected photoelectrons of i_{th} fired PMT and the distance between i_{th} PMT and shower core.
- Density ratio, DR, is defined as: $DR = \frac{\sum PE_{50} / \sum PMT_{50}}{\sum PE_{10} / \sum PMT_{10}}$. It is the ratio of average density at two radial distance, 50 m and 10 m.
- r70, minimum radius which contained 70% PEs of the whole shower.
- Pincness, is defined as: $P = \frac{1}{N} \sum_{i=0}^{N} \frac{(\zeta_i - \langle \zeta_i \rangle)^2}{\sigma^2_{\zeta_i}}$. It describes the "clumpiness" of the lateral distribution of an event. Where $\zeta_i = \log_{10}(PE_i)$, for each hit, an expectation $\langle \zeta_i \rangle$ is assigned averaging the ζ_i in all PMTs contained in an annular containing the hit, with a width of 5 m, centered at the core of the air shower [12].
- PMT Ratio, PR. It's the proportion of chosen PMTs within all fired PMTs. If the charge of one PMT is at least three times of the averaging charge in an annular containing this PMT with a width of 5 m, it will be one of the chosen PMTs.

4. Multivariate analysis

There are many multivariate data analysis algorithms now. For the scientific community, most of these nonlinear methods can be found in Toolkit for Multivariate Analysis(TMVA) package, which provides a ROOT-integrated environment for the processing, application of multivariate classification. The TMVA methods are including but not limited to K-Nearest Neighbour(K-NN) Classifier, H-matrix discriminant, Article Neural Network(ANN), Boosted Decision Trees with a Gradient boosting algorithm(BDTG), Support Vector Mathine(SVM). In this article, two kinds of the method are used for gamma/hadron separation, Multilayer Perceptron Article Neural Network(MLPANN) and the BDTG. ANN is a computational or mathematical model, based on bio-



(a) MLPANN (b) BDTG

Figure 3: MLPANN and BDTG output for gamma (blue line) and background (red line) events

logical neural networks, and a delicate method in multi-variate analysis algorithms[9, 10]. Basically, it has three different layers(input layer, hidden layer, and output layer) and seven different types(multilayer perceptron, convolutional neural network and so on) now. A decision tree takes a set of input features and splits input data recursively based on those features, shows good result in non-linear dependencies. Gradient boosting is known to be one of the leading weighted ensemble algorithms. The BDTG increases the performance of classifier and stabilizes the response of the decision trees with respect to fluctuations in the training sample [11]. In this work, we put all the seven sensitive parameters into multivariate training and testing. Example output histogram for the test sample with MLPANN and BDTG methods in the multiplicity bin $200 < nfitc < 300$ are shown in Fig 3.

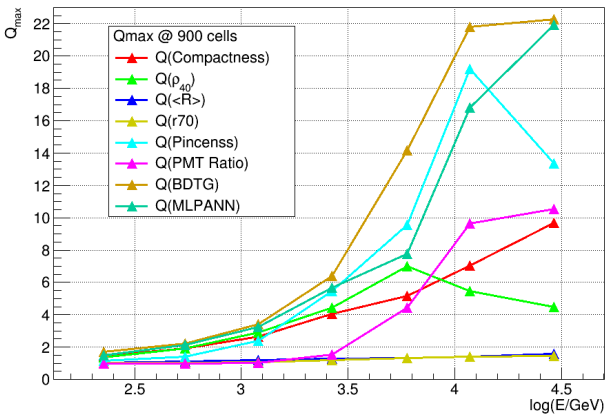


Figure 4: Q_{max} from different methods

5. Separation results of different methods

Every data sample has been grouped into 7 bins according to the number of nfitc of an event, where nfitc means the number of PMTs during curvature fitting. The performance of data classification can be evaluated by the quality factor Q. It is defined as $Q = \varepsilon_\gamma/\sqrt{\varepsilon_p}$, where ε_γ is the fraction of gamma events selected by the cut and ε_p is the fraction of proton events passing the cut. To keep enough gamma events in analysis, $\varepsilon_\gamma > 0.5$ is used in calculating the Q factor. Fig shows the final separation results of these single parameter methods, by optimizing the cut value of parameters Q_{max} can achieve a very high value in the high energy bin. At the same time, we can tell that parameter C, ρ_{40} have shown a better performance during single parameter discrimination.

To get better separation in low energy bins, multivariate analysis methods are applied with all sensitive parameters. For each multiplicity bin, half of the data was used for training, the other half for validation. In each methods, a cut value on the normalized output was selected by finding the maximal quality factor in the training phase of the analysis. However, the multivariate methods didn't show its advantage over single parameter method. Further study should be done to find out the reason. Fig 4 shows the detail maximum Q factor get from different methods with different energy. The quality factor Q get smaller at high energy because of the inadequate volume of simulation data.

6. Discussion

Since April 2019, one pool of WCDA has been started test running and taking data and some very preliminary analysis of some famous sources has been done. Taking the Compactness as an example, the comparison between MC data and early experiment data has been processed. Such as the fraction of background hadron events passing photon/hadron discrimination cut, the distribution of cut parameters in every multiplicity bin and so on. They can match pretty well at certain bins. Almost 99% of the background events can be rejected after the Compactness cut with energy about 3 TeV.

We have a very preliminary result of significance obtained on the source from the Crab after the gamma/hadron discrimination the significance of crab has a obviously improve. The results indicate the selection is very effective. However, there are still some problems need to be solved which requires further investigations. Better results will be induced by a better understanding of experiment data.

7. Conclusion

This study has demonstrated that LHAASO-WCDA has an excellent gamma/hadron separation power, in other words, an excellent background rejection power. The multivariate analysis methods with four sensitive parameters can significantly improve the gamma/hadron discrimination performance at low energy range. During 90 days of running, experiments data still can't match simulation data perfectly, further study should go on to get a better detector performance.

Acknowledgments

This work is supported in China by NSFC (No.2018YFA0404201,U1831208,11675187), the Chinese Ministry of Science and Technology, the Chinese Academy of Sciences, the Key Laboratory of Particle Astrophysics, IHEP, CAS.

References

[1] Z.cao et al, Proceedings of 33th ICRC(2013).

[2] http://www-ik.fzk.de/corsika

[3] S. Ostapchenko, Phys. Rev. D, 74 (1)(2006): 014026.

[4] J.R. Horandel, Astroparticle Physics, 19 (2003): 193220.

[5] S. Agostinelli, et al, Geant4 Collaboration, Nucl. Instrum. Methods Phys. Res. A,506(2003):250303.

[6] http://neutrino.phys.ksu.edu/GLG4sim.

[7] H.C. Li, et al, Chin. Phys. C, 41 (2)(2017): 026002.

[8] Min zha et al,,Comparison of measured and simulated data with early LHAASO-WCDA run data,Proceedings of 36th ICRC(2019).

[9] R.K. Bock, et al. NIM A 588 (2008) 424.

[10] M.D Richard, L. Lippmann, Neural Computation 3(1991) 461.

[11] J.R. Quinlan, et al., In Proceedings of the Thirteenth National Conference on Artificial Intelligence,(1996).

[12] A.U. Abeysekara et al., ApJ 843 (2017), 39.