PROCEEDINGS OF SCIENCE

Primary particle identification with MVA method for the LHAASO project

Zizhao Zong[a]*, Baiyang Bi[b], Lingling Ma[b], Liqiao Yin[a], Shoushan Zhang[b], Tiina Suomijärvi[a], and Zhen Cao[b] for the LHAASO collaboration

[a] *Institut de Physique Nucléaire d'Orsay, IN2P3-CNRS, Université Paris-Sud, Université Paris-Saclay, 91406 Orsay Cedex, France*

[b] *Institute of High Energy Physics, Chinese Academy of Sciences, Beijing 100049, China*

The LHAASO (Large High Altitude Air Shower Observatory) project, which is under construction at high altitude of 4400m a.s.l. in Sichuan, China, aims to observe the extensive air showers (EAS) induced by cosmic rays in the atmosphere. LHAASO consists of several large detector arrays including KM2A (1 km^2 array), WCDA (Water Cherenkov Detector Array) and WFCTA (Wide Field of view Cherenkov Telescope Array). By employing hybrid detection technique, LHAASO offers an accurate measurement of the cosmic-ray spectrum and composition around the knee region. Furthermore, the primary particle identification can be obtained by using Multivariate Analysis (MVA). In this contribution, we present the parameters that will be measured by various detectors of LHAASO in the EAS detection and discuss the performance of the MVA method for primary particle identification.

35th International Cosmic Ray Conference — ICRC2017
10–20 July, 2017
Bexco, Busan, Korea

*Presenter

PoS(ICRC2017)547

1. Introduction

The LHAASO project, located at 4400 m a.s.l. in Sichuan Province, China, is a multi-purpose project for the detection of high-energy gamma rays and cosmic rays with hybrid techniques. LHAASO is expected to solve some open questions in Galactic cosmic-ray physics by studying the extensive air showers (EAS) induced by both charged particles and gamma rays. The LHAASO observatory covers an area of 1 km^2 and consists of three detector arrays, the 1 km array (KM2A), the Water Cherenkov Detector Array (WCDA) and the Wide Field of view Cherenkov/fluorescence Telescope Array (WFCTA). Currently, the Observatory is under construction. One fourth of the Observatory is expected to be finished in the year of 2018 and the whole Observatory is expected to be completed by the end of 2021.

LHAASO proposes to measure the energy spectrum and identify the mass composition of cosmic rays around the "knee" region, the origin of which is still under discussion. In the reconstruction of air shower events, there is a strong mutual dependency between the primary energy and the primary particle type. The air showers with primary energies around PeV level just reach the maximum of shower development around the observatory level of 4400 m a.s.l., yielding minimal shower fluctuations. The combination of different detector arrays offers a large amount of data, including the parameters related to shower size, shower geometry, and muonic component, for the shower reconstruction. Benefiting from the advantages of the high altitude of the site and the hybrid observation of the EAS, LHAASO is capable of measuring the energy and classifying the chemical nature of the primary particles with high accuracy [1].

The simulation of each LHAASO detector array has been developed [2, 3, 4]. Various parameters, correlated to the shower properties such as the primary energy, arrival direction, and primary particle, can be determined from the simulated data collection. In this work, we applied the MVA method, integrated in the TMVA package [5], for primary particle identification based on the simulation of the LHAASO hybrid detectors. We will present the parameterization of data that will be measured by each detector array and the result of the MVA classification.

2. Parameterization of data from LHAASO hybrid detectors

The schematic of LHAASO Hybrid detectors is shown in Fig.1. The WCDA is located in the central area of the observatory, covering a 300 m $\times$ 260 m rectangular area. It consists of three large water pools, which are segmented into 5 m $\times$ 5 m grids as single detector units. WCDA has a large duty cycle (>90%) and a wide field of view ($\sim$1.5 sr) in the detection of EAS. In cosmic-ray measurements, one of the WCDA pools (150 $\times$ 150 m^2) is used for the detection of shower geometries. Each cell in this pond is equipped additionally with a 1-inch small photomultiplier tube (PMT), which can extend the dynamic range for cosmic-ray detection [7]. The telescopes of WFCTA are deployed alongside this shower core detector array providing calorimetric energy measurements. Each telescope has a field of view (FoV) of $14^\circ \times 16^\circ$. The KM2A, surrounding the central area, is a complex array composed of electromagnetic particle detectors (ED) and muon detectors (MD). The KM2A detectors are uniformly distributed in the remaining area of the observatory, covering nearly 1 km^2. In the EAS detection, these detector arrays can be combined together for hybrid measurements, yielding a precise reconstruction of the shower parameters.

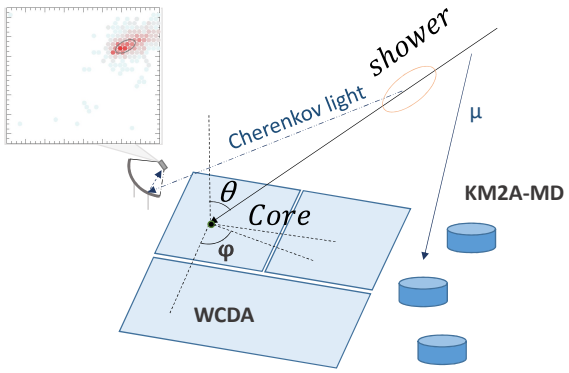

Figure 1: Hybrid detection of the EAS at the LHAASO Observatory

A shower library simulated by using the CORSIKA package [6] with a slope of -2.3 for the energy spectrum over the energy range from 100 TeV to 1 PeV, covering a FoV from 22° to 38° for zenith (θ) and from 77° to 103° for azimuth (φ), is used in this work. The shower library consists of five mass compositions for the cosmic rays: p, He, CNO, MgAlSi and Iron. The responses of the three detector arrays are simulated for the shower parameters saved in binary files. In the detector simulation, one WFCTA telescope, pointing to $(\theta, \varphi) = (30°, 90°)$ is employed to work together with WCDA and KM2A-MD. The shower events triggered by both the WCDA-core detector array and the WFCTA telescope are selected for further analysis.

Parameterization of the WCDA data Fig.2 shows the signals in p.e. numbers measured in each cell of WCDA for a shower event. The core position and the arrival direction of the shower can be obtained with a precision of ∼4.2 m by fitting the hump of S_{WCDA}. Fig. 2 shows the lateral distribution of S_{WCDA} (binned in each 5 m). By fitting the lateral distribution of S_{WCDA} with an NKG-type function (Eq.2.1), the reference parameter of WCDA signals, S_{ref}, can be obtained to describe the shower size. In Eq.2.1, R_{ref} is set to 700 m and β and γ are variables related to the shower geometry and shower size.

$$S(r) = S_{ref} \cdot \left(\frac{r}{R_{ref}}\right)^{\beta} \left(1 + \frac{r}{R_{ref}}\right)^{\beta+\gamma} \tag{2.1}$$

Parameterization of the KM2A-MD data The muon detector (MD) in KM2A consists of water Cherenkov detectors, which have been proved to have high stability and to be cost-effective. Each MD is a cylinder-shape water tank with a diameter of ∼6.8 m and a height of ∼1.2 m. MDs are buried in the ground with a 2.5 m-thick layer of soil shielding low-energy electromagnetic components of the air showers. Therefore, the responses of MDs in EAS detection are mostly induced by the muon components, which are of high importance for determining the primary mass and charge of the EAS.

The complete information of KM2A simulation is introduced in Ref. [2]. As a result of the shower detection, the number of muons, n_μ, measured in each MD for a certain event is available for further analysis. Fig.3-left shows the distribution of n_μ for a given shower. The binned n_μ

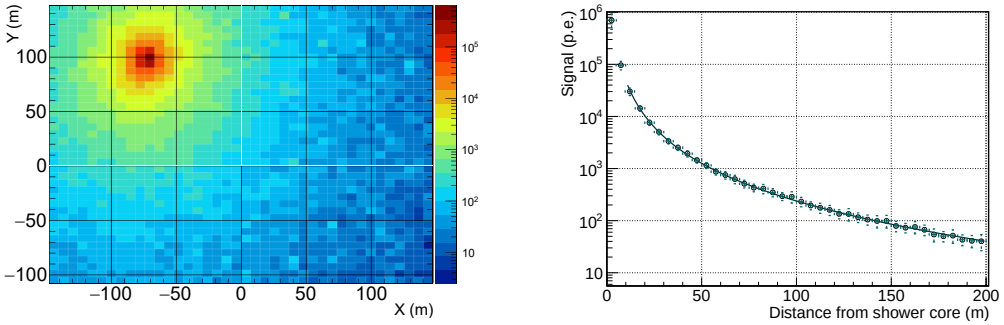

Figure 2: *Left:* 2-D distribution of S_{WCDA} from each detector cell in number of photoelectrons. *Right:* Lateral distribution of a given shower measured by WCDA.

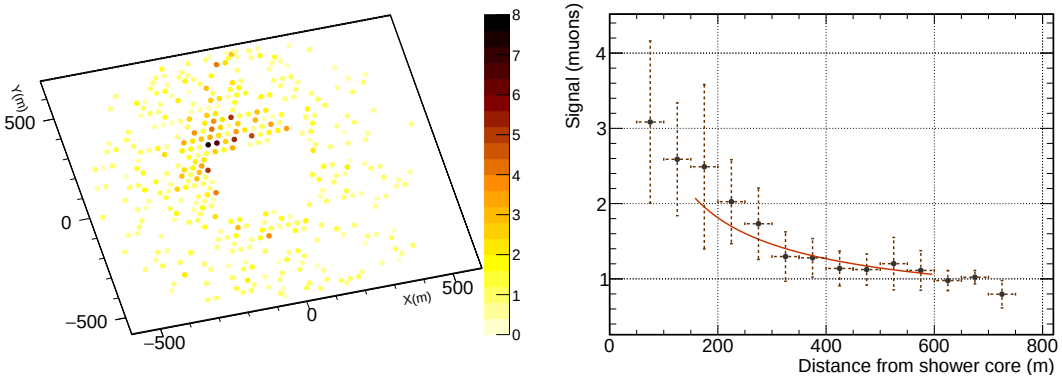

Figure 3: *Left:* 2-D distribution of muon numbers measured by KM2A-MD. *Right:* Lateral distribution of muon shower measured by KM2A-MD.

of each 50 m with respect to the core distance is shown in Fig.3-right. Similarly to the case for S_{WCDA}, a reference of muon number, n_μ^{ref} can be obtained by fitting the n_μ with a lateral distribution function of a muonic shower (Eq.2.2) [8], where R_{ref} is set to 400 m and η depends on the zenith angle and the primary energy of the shower.

$$n_\mu(r) = n_\mu^{ref} \cdot \left(\frac{r}{R_{ref}}\right)^{-\frac{3}{4}} \left(\frac{r+320}{R_{ref}+320}\right)^{-\eta} \tag{2.2}$$

Parameterization of the WFCTA image The response of the WFCTA telescopes to the shower event is generally an elliptic image in the camera (see Fig.4). Each image is first cleaned from the NSB (night sky background) noise by removing pixels with less than one neighbor and pixels below 10 p.e.. After the cleaning, we use the Hillas parameters, which has been widely employed in the experiments with Cherenkov telescopes [9], to describe the image.

The *SIZE*, integrated p.e number of the image, is a crucial parameter correlated to the energy and the core distance from the telescope. The other parameters such as *Width*, *Length*, *Dist* and *Miss* are correlated to the shower geometry and the longitudinal development of the shower (See Fig.4). The WCDA has good accuracy for the reconstruction of shower cores (∼4.2 m) and arrival

directions ($\sim$0.3°). By combining the core position measured by the WCDA and the image parameters from WFCTA, the energy of the primary particle can be determined with a precision of $\sim 20\%$.

3. Primary particle identification with MVA method

As the multivariate methods have significantly evolved in the recent years, they have become a potential classification tool for most data analyses in high-energy physics and astroparticle experiments. Compared to traditional cut-based analysis techniques such as liner-cut classification, likelihood classification, or Fisher discriminants, MVA methods such as Neural Networks (NNs) or Boosted Decision Trees (BDT) have several advantages. The main strength of them is the consideration of non-linear correlations between input parameters, which is crucial for the analysis of a complex data set with multidimensional information. After a comparison between series of BDT and NNs classifiers provided by the TMVA package, the Boosted Decision Trees (BDT) method is selected to be employed in this work, since it has an obvious advantage in the separation of compositions and it is also faster for algorithm training than the NNs classifiers.

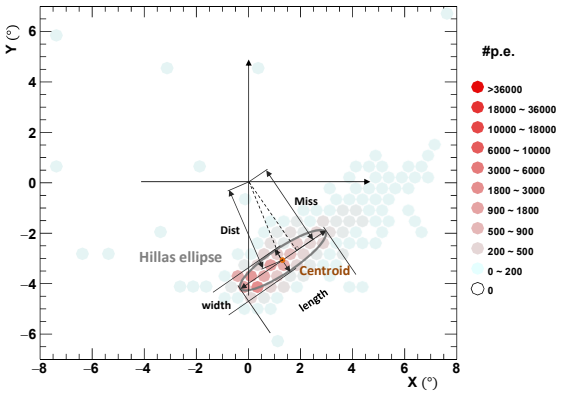

Figure 4: WFCTA image and Hillas parameterization

3.1 Parameters for particle identification

Before training the MVA algorithms, a tuning of parameters is implemented. Four parameters, P_{lat}, P_{μ}, P_E and P_{long}, are finally used as input parameters for the MVA classifiers (see Eq.3.1, Eq.3.2, Eq.3.3, Eq.3.4).

P_{lat} is expressed as Eq.3.1,

$$P_{lat} = log_{10}\left(\frac{S_{max}}{S_{ref}}\right) \tag{3.1}$$

where S_{max} is the maximum signal among WCDA cells, and S_{ref} is the shower size measured by WCDA. This parameter is strongly correlated to the residual energy of shower at the ground level and the X_{max} of the shower.

The expression of P_{μ} is given by Eq.3.2

$$P_{\mu} = log_{10}(\, N_{\mu} + n_{\mu}^{ref}\,) + log_{10}(\, N_{MD}\cdot n_{\mu}^{ref} + n_{\mu}^{ref}\,) \tag{3.2}$$

where N_{mu} is the total muon number measured by the MDs located in the range of core distance from 150 m to 600 m. As the shower core is measured by the WCDA core detector, a fair proportion of muons arriving at the area close to the core can not be measured by MDs due to the layout of KM2A. Therefore, MDs near the shower core (<150 m) are removed from the counting of total muon number to unify the rule of parameterization and to reduce the bias induced by the detector layout. The upper limit is set to 600 m to reduce the bias due to the incomplete counting for

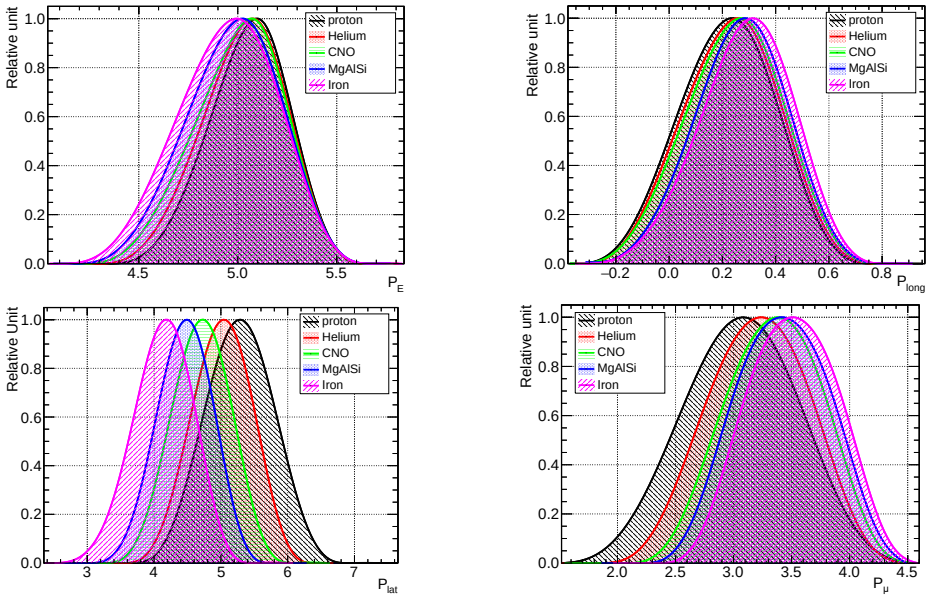

Figure 5: Parameter distributions for different mass compositions

the muons arriving at the edge of KM2A. n_μ^{ref} is the fitted parameter from Eq.2.2 and N_{MD} is the number of triggered MDs for a given shower. P_μ is a crucial parameter for particle identification, as the produced muons in the EAS are only affected by the ionization when propagating in the atmosphere and basically arrive at the ground with a very low degradation of energy.

P_E is expressed as a logarithmic form of $SIZE$ corrected by R_p and given by Eq.3.3,

$$P_E = log_{10}(SIZE) + 0.0084 \cdot R_p \tag{3.3}$$

where R_p is the distance from the telescope to the shower axis. P_E is the main parameter correlated to the energy of shower.

The expression of P_{long} is given by Eq.3.4,

$$P_{long} = \frac{Width}{Length + Dist + Miss} - 0.003 \cdot R_p \tag{3.4}$$

which is composed of the geometry-related parameters ($Width$, $Length$, $Dist$, and $Miss$) from the WFCTA image and also corrected by R_p. P_{long} is correlated to the X_{max} of the shower.

Fig.5 shows the distribution of these four input parameters for all the selected shower samples initiated by different mass compositions of cosmic rays.

3.2 Training and application of the BDTG classifier

Decision trees have a structure of nodes, which determine the response to a given event based on the logic that is built during the algorithm training. In the classification, a set of parameters for a given event is sent to each node for a binary decision, until the final decision of *Signal* or *Background* is made for this event.

The method of Boost Decision Trees with Gradient boosting (BDTG), provided by the TMVA package, is employed as classifier in this work. In the algorithm training of the BDTG classifier,

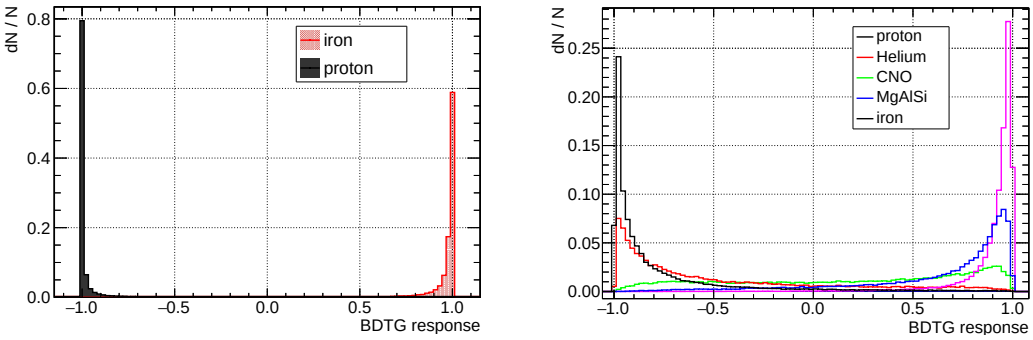

Figure 6: Responses of different compositions in the Iron/p (*left*) and heavy nuclide/(p+He) (*right*) separation with BDTG classifiers for all the selected event.

most parameters of training correspond to the default settings, since they have been tested and set to the optimal values by the TMVA team. Some parameters are modified for a balance between the performance of the classifier and the time consumption by the processors or to avoid overtraining.

- The number of decision trees is adjusted to 500 (default value: 1000). The tests with various values from 500 to 1000 show that it does not significantly affect the current result.
- The number of the grid points in variable range, used to select the optimal cut value in the binary decision at each node, is changed to 50 (default value: 20) to sufficiently optimize the performance of the classifier.
- The weight of each event in the training samples is defined as $\frac{1}{(dN(E)/N)\cdot N}$, where N is the number of events for each composition and $dN(E)/N$ is the relative flux for a given event with a primary energy of E.

3.3 Results and discussion

The BDTG classifiers for the separation of Iron/p and heavy nuclide/(p+He) are trained and applied to the simulation data of LHAASO hybrid detection. The responses of the two separations are shown in Fig.6-left and Fig.6-right. The cut values for Iron/p and heavy nuclide/(p+He) separations are set to 0 and to -0.1, respectively, based on the analysis of *Signal-to-Background* ratio. The cut efficiency and the contamination of each separation are calculated based on the response of each event weighted by its proportion of flux in the cosmic-ray spectrum following the Höerandel model [10] (see in Fig.7). For the separation of iron/p, nearly perfect results approximating the theoretical limit are obtained with the cut efficiency $> 95\%$ and the contamination $< 5\%$ over the energy range from 100 TeV to 1 PeV. For the separation of heavy nuclide/(p+He), a clear separation for the different responses is obtained from the BDTG classi-

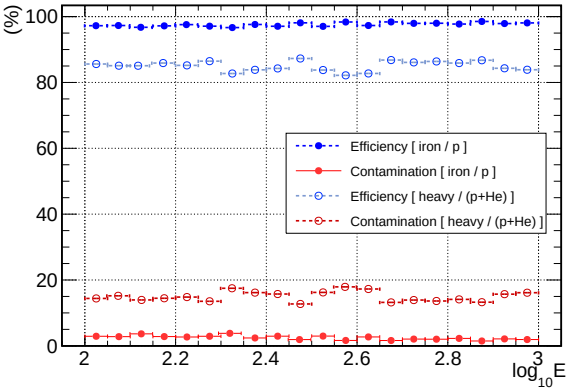

Figure 7: Efficiency and contamination of the heavy/(p+He) and iron/p separations.

fication. The cut efficiency and the contamination are $\sim 85\%$ and $\sim 15\%$ over the given energy range, respectively.

The performance of classifiers highly depends on the separation of different mass compositions with each variable. According to the variable ranking by the TMVA and the classifiers, P_μ is the most important variable for the classification, which is expected based on the physics mechanism of the EAS development. P_{lat} ranks the second as it's sensitive to X_{max} of the showers. P_E and P_{long} are both parameters which are tuned from the Cherenkov image sampled with a single telescope at the observatory level. They are specially used for energy reconstruction of the showers and much less efficient than P_μ and P_{lat} for the separation with individual variables. However, they are also helpful in a fair proportion of the decision trees, since the shower energy and the primary mass are interrelated in the classification.

4. Summary

Primary particle identification based on the simulations of the LHAASO hybrid detection of the EAS has been implemented by using MVA methods over the energy range from 100 TeV to 1 PeV. Simulated data for various detector arrays of LHAASO are parameterized and tuned for the training of the MVA classifiers. The first results show perfect separations for iron/p and good separations for heavy nuclide/(p+He) with the BDTG classifier. Further studies will be performed with better statistics for mass composition and for a larger energy range of cosmic rays.

References

[1] L. Q. Yin, et al. for the LHAASO Collaboration, *Accurate Measurement of the Cosmic Ray Proton Spectrum from 100TeV to 10PeV with LHAASO*, these proceedings.

[2] S. W. Cui, et al. *Simulation on gamma-ray astronomy research with LHAASO-KM2A*, Astropart. Phys. 54 (2014): 86-92.

[3] Z. G. Yao, et al. for the LHAASO Collaboration, *LHAASO Simulation: Performance of the Water Cherenkov Detector Array*, Proc. Int. Cosmic Ray Conf. 3, Łódź, Poland, 2009.

[4] S. S. Zhang, et al. *Properties and performance of two wide field of view Cherenkov/fluorescence telescope array prototypes*, Nucl. Instrum. Meth. A 629.1 (2011): 57-65.

[5] A. Hoecker, et al. *TMVA-Toolkit for multivariate data analysis*, arXiv preprint physics/0703039 (2007).

[6] D. Heck, et al. *CORSIKA: A Monte Carlo code to simulate extensive air showers*, No. FZKA-6019. 1998.

[7] C. Liu, et al. for the LHAASO Collaboration, *The dynamic range extension system for the LHAASO-WCDA experiment*, these proceedings.

[8] J. G. Gonzalez, *Measuring the muon content of air showers with IceTop*, EPJ Web of Conferences. Vol. 99. EDP Sciences, 2015.

[9] A. M. Hillas, *Cerenkov light images of EAS produced by primary gamma*, roc. Int. Cosmic Ray Conf. 3, La Jolla, United States, 1985.

[10] J.R. Höerandel, *On the knee in the energy spectrum of cosmic rays*, Astropart. Phys. 19.2 (2003): 193-220.