32ND INTERNATIONAL COSMIC RAY CONFERENCE, BEIJING 2011

Gamma/hadron separation study for the HAWC detector on the basis of the multidimensional feature space using non parametric approach

V. GRABSKI[1], L. NELLEN[2], A. CHILINGARIAN[3], A. VARDANIAN[3], FOR A COMPLETE AUTHOR LIST, SEE THE SPECIAL SECTION OF THESE PROCEEDINGS
[1]*Instituto de Fsica, UNAM, México*
[2]*Instituto de Ciencias Nucleares, UNAM*
[3]*Alikhanyan National Science Laboratory, Armenia*
grabski@fisica.unam.mx

DOI: 10.7529/ICRC2011/V09/1191

Abstract: The High Altitude Water Cherenkov (HAWC) is a high duty cycle, wide field of view gamma-ray observatory that will be located in Mexico at an elevation of $4100\,\mathrm{m}$ (lat. 19° and long. 97°). The gamma/hadron separation study is performed on the base of the non parametric methods for different probability density estimators, using a multidimensional feature space. Data from full detector simulations of the HAWC observatory are used to estimate the features. The gamma/hadron separation study is performed using different methods of multivariate analysis. A study of sensitive parameters for gamma/hadron separation versus signal efficiency, obtained by the use of different cuts on the classifier outputs, is presented. The results show a significant reduction of the background contamination in the region of high signal efficiency uaing multivariate analysis, compared to the default procedure proposed for HAWC.

Keywords: HAWC, gamma ray observatory, hadron rejection, neural networks

1 Introduction

The gamma/hadron separation for the low energies ($<$ $1\,\mathrm{TeV}$) is difficult task for ground base gamma observatories. In this case only a small part of shower reaches to the ground and the difference between the topology of hadronic and electromagnetic showers can not be used effectively. The detection of muons is another signature to separate the gamma and hadron showers, but for low energies muon detection does not allow better discrimination. The features proposed here should effectively extend both signatures.

Recently different features have been developed [1–3] to distinguish between the photon and hadron showers using multivariate analysis. The choice of features depends on the detector performance, so they should be optimized for each individual detector. In this work, we present some choice of the features and their efficiency for the HAWC detector. The HAWC [4] detector design considers 300 cylindrical water tanks of $7.3\,\mathrm{m}$ diameter and $4.5\,\mathrm{m}$ height, with three Hamamatsu PMTs (R5912) at the bottom.

1.1 The choice of features

In difference from scintillator based EAS arrays, which sample the charged particle density at different distances from the shower core, HAWC detects the energy deposition that can be considered somehow correlated with the particle density for the electromagnetic component of the shower. Muons significantly affect that correlation and they should appear as local maximums in the image. Therefore, it is possible to use muon signatures along with the density behavior of showers for gamma/hadron discrimination.

The results here have been obtained using the full simulation of HAWC detector. The sample of events used in this work has been generated by means of the Monte Carlo simulation program CORSIKA 6.7 [5]: $\approx 10^9$ gamma and proton events have been simulated within the energy range $[0.05, 100]\,\mathrm{TeV}$, with the primary energy spectrum following a power law with spectral index -2.0 for all primaries, with zenith angle between 0° and 75°. The shower cores are placed on a circle of radius $1000\,\mathrm{m}$ around the center of the array. The detector response has been simulated using g4sim, a tool developed by the Milagro and HAWC collaborations, which is based on the GEANT4 package [6]. The analysis has been performed in the shower plane, using the arrival direction and the core position reconstructed by the HAWC collaboration's analysis program AERIE. Only the events with a successful angular reconstruction have been analyzed.

The feature vector is constructed using some of the existing features for gamma/hadron discrimination in the literature [3] and some new elements obtained using HAWC detector full simulation data. Initially, a 6 dimensional fea-

ture vector was under consideration. These features should more or less reflect the topology difference in gamma and hadron showers as well as the muon richness of the hadron showers.

• $HAWC_{\rm gh}$ (the standard algorithm) is defined as: $N_{\rm hits}/PE_{\rm max}(> R)$, where $N_{\rm hits}$ is the hit multiplicity and $PE_{\rm max}$ is the maximum number of photo electrons detected outside the radius R from the core. The detected muons contribute a large signal density, which will increase the power of this feature for the gamma/hadron discrimination.

• mean radius RM is defined as: $\sum(PE_iR_i)/\sum PE_i$, where PE_i is the number of photoelectrons in hit i and R_i is the distance of hit i from the core. This variable should be different for gamma and hadron showers.

• PER feature is defined as: $N_{\rm hits}/\sum(PE_iR_i)(> R \text{ and } PE_i > Cut)$, where the sum is calculated outside the radius of R from the core center including only PE_i values that are larger than some cut value. The divisor of this variable is associated with the the integral of the radial density where the hadron showers dominate gamma showers. The use of the cut on PE's in the sum is to enhance the muon contribution to this feature.

• $PEMRM$ feature is defined as: $N_{\rm hits}/(PE_{\rm max}R_{\rm max})$, where $R_{\rm max}$ is the position of the hit with the maximum PE value outside the radius R from the core center.

Taking into account that all features have larger values for gamma showers than for hadronic ones leads us to expect that a new feature obtained by the multiplication of any pair will have better discrimination power than the each feature alone. The next two features can be considered as complementary ones.

$$\begin{aligned} P1 &= HAWC_{\rm gh} \times PER \\ P2 &= PER \times PEMRM \end{aligned}$$

The $HAWC_{\rm gh}$ feature is already in use in the official HAWC software as the single gamma/hadron discrimination parameter with some initial optimization already performed. Particularly, it has been found that the optimal value of the parameter R is $40\,\mathrm{m}$ (denoted $HAWC40_{\rm gh}$). In this study, the optimization processes of this feature will be continued. The optimization of all features is done for individual features as well as combining several features.

From the above mentioned, it can be noticed that all variables depend on the estimation of the core position. So the optimization of the R for each feature depends also on the algorithm for determining the core position. Furthermore, it should be mentioned that the precision of the core position estimate is different for the gamma and hadron showers as well as for the showers having the core position inside and outside the detector area. It is, therefore, imporant to start determining the optimal method for estimating the core position in order to get better discriminating power. Three methods have been used for the core position estimation. Two of them are already implemented in AERIE (the barycenter and the Gaussian fit to the lateral distribution [7]) and third one is the position of the PMT having the maximum PE value. The precision of the core position estimation depends on the hit multiplicity. This means the parameter R should be optimized separately for different regions of hit multiplicity.

Mult. Wind.	10–100	100–400	> 400
PER	$10\,\mathrm{m}$	$25\,\mathrm{m}$	$40\,\mathrm{m}$
$HAWC_{\rm gh}$	$20\,\mathrm{m}$	$45\,\mathrm{m}$	$60\,\mathrm{m}$
$PEMRM$	$30\,\mathrm{m}$	$40\,\mathrm{m}$	$60\,\mathrm{m}$
P_1	$10\,\mathrm{m}$	$30\,\mathrm{m}$	$50\,\mathrm{m}$
P_2	$15\,\mathrm{m}$	$30\,\mathrm{m}$	$50\,\mathrm{m}$

Table 2: The optimal values of R for all features and for different hit multiplicity windows.

Initially the optimal values of R for each variable and for the different intervals of hit multiplicity have been obtained by the maximum of two parameters characterizing the separation of the distributions of two classes (gamma and hadron). The first is the well known Students t parameter defined as:

$$t = (m_g - m_p)/\sqrt{s_g^2 + s_p^2}$$

where m_g, s_g and m_p, s_p are the mean values and the standard deviations of the given feature distribution for the gamma and proton showers respectively. The second is similar to the t parameter and is defined to measure some relations between cumulative probability functions of two classes:

$$d = (G^{-1}(0.17) - P^{-1}(0.83))/(G^{-1}(0.5) - P^{-1}(0.5))$$

Where G and P are the corresponding cumulative distributions for gammas and protons, respectively. If the d parameter value is positive then we have at least 83% background rejection with the signal efficiency 83%. These exclusive values have been chosen to compare $d = 0$ to $t = 1$. For a Gaussian distribution these two conditions are realized for the same variable value.

The results of both parameters are presented in table 1 for different variables, using events with more than 400 hits. One can notice that R value that maximizes the discrimination using $HAWC_{\rm gh}$ feature is about $60\,\mathrm{m}$. The obtained optimal values of parameter R for the three hit multiplicity windows and for all features are presented in table 2. These two parameters also have been used to pick the core position estimation method which is most suitable for this task. The feature $HAWC40_{\rm gh}$ in table 1 is the one currently used in AERIE, along with the Gaussian core fit. The result is the same using the maximum PE position as a core center. One can notice a slight improvement in the d and t parameters for the $HAWC_{\rm gh}$. Comparing the columns 30bc and 30pem of table 1 a small improvement in d and t is noticed in case of using the maximum PE position as the core center compared to using the BC (barycenter).

R	10 m		20 m		30 m (pem)		30 m (bc)		40 m		50 m		60 m		70 m	
	t	d	t	d	t	d	t	d	t	d	t	d	t	d	t	d
PER	0.75	-0.8	0.82	-0.67	0.86	-0.57	0.82	-0.67	0.84	-0.5	0.7	-0.8	0.52	-1.4	0.6	-0.86
$HAWC_{\rm gh}$	0.51	-1.3	0.78	-0.28	1.07	-0.01	1.14	0.03	1.27	0.1	1.38	0.15	1.38	0.18	1.2	0.17
$PEMRM$	0.81	-0.1	1.02	0.06	1.21	0.12	1.22	0.06	1.37	0.18	1.37	0.18	1.33	0.18	1.2	0.15
P_1	0.49	-0.8	0.68	-0.27	0.88	-0.09	0.88	-0.12	0.98	-0.02	0.93	0.01	0.8	-0.03	0.72	-0.08
P_2	0.65	-0.2	0.79	-0.07	0.92	-0.01	0.94	-0.06	0.96	0.02	0.9	0.01	0.76	-0.01	0.58	-0.09
$HAWC40_{\rm gh}$									1.2	0.06						

Table 1: Results of t and d parameters dependent on R for all features for $N_{\rm hit} > 400$.

	Event count		Median E [TeV]	
	Gammas	Protons	Gammas	Protons
10–50	120063	355434	0.38	0.8
50–100	68440	128279	0.9	1.7
100–200	47508	86145	1.6	2.7
200–400	30351	56805	3.0	4.6
> 400	23581	49553	11	15

Table 3: Number and median energy of analyzed events in each multilicity window

Mult. Wind.	10–50	50–100	100–200	200–400	> 400
PER	1	1	1	3	6
$HAWC_{\rm gh}$	2	5	5	4	2
$MPEMR$	6	4	4	5	1
P_1	3	3	3	1	5
P_2	7	2	2	2	4
$HAWC40_{\rm gh}$	4	6	6	6	3
RM	5	7	7	7	7

Table 4: MLP_ANN Variable Ranking Table

2 Multivariate algorithms and results

There are many multivariate data analysis algorithms described in the literature. Which one to pick depends on the volume of the statistics, feature space dimensionality and the time for the classification of an event. Most of these algorithms can be found in the MVA [8] program package which is available for the scientific community and can be easily used inside the ROOT program [9]. Many nonlinear methods of TMVA like Artificial Neural Networks (ANN nonlinear discriminant analysis), Boosted Decision Trees, Support Vector Machine, as well as the ANI package [10] ANN. are good enough for this task and yield a similar performance. Linear discriminant methods like Fisher and some other modifications are faster than the ones mentioned above, but are slightly inferior in performance and less applicable if the features are correlated, as is the case here, and require an initial decorrelation procedure. In this study, we present only ANN in order not to complicate the picture.

Three neural network implementations are available in TMVA. The MLP ANN has been used, being a fast and flexible implementation, recommended by TMVA. The number of the hidden layers in the network and the number of neurons in these layers are configurable. If the available computing power and the size of the training data sample is sufficient, one can increase the number of neurons in the hidden layer until the optimal performance is reached. The possibility for the variable ranking exists which allows to reduce the size of the feature vector, removing less important inputs.

To perform the data classification, the events have been grouped in 5 multiplicity bins. In this case more windows have been used since the gamma/hadron discrimination is more sensitive to hit multiplicity. For each multiplicity window, training and classification procedures have been performed using MLP ANN. Half of the data in table 3 was used for training, the other half for validation. The neural net used has one hidden layer. The optimal number of neurons in the hidden layer was found to be $N_{\rm variables} + 5$. The results of the ANN classification have been presented in two ways: the signal (gamma) efficiency dependent on the background(proton) rejection efficiency and the signal efficiency dependent on the quality factor Qf [3]. For the quality factor Qf we use the definition given in study [3]

$$Qf = e_\gamma / \sqrt{1 - e_{\rm bkg}}$$

where e_γ is the fraction of gamma-ray showers selected by the cut, and $1 - e_{\rm bkg}$ is the fraction of hadrons misidentified as photons. To start, the discrimination power of each feature has been studied separately for each multiplicity window. In table 4 we present the feature rankings by MLP ANN for each multiplicity window. These results are consistent with the results of Students t-test and the values of d parameter. One can notice that the PER feature is the best one in three hit multiplicity windows without the cuts on PE values.

In fig. 1, we show the signal efficiency versus the rejection factor for all the features. Almost all the features have the same discrimination power, except in the high signal efficiency region. The increase there is mostly due to the feature PER. The background contamination at 80% signal efficiency is reduced by about 60% when all features are applied together in the classification procedure (compared to the use of a single feature). In fig. 2, the dependence of Qf on the signal efficiency for the combined features is shown. The Qf parameter gets its maximum value 2.5 at the signal efficiency 80% and the improvement of the multiple feature analysis compared with a single feature case is about a factor of 1.2 at 80% signal efficiency for the hit multiplicity window 100–200.

Fig. 3 presents the same information as fig. 2 but for different multiplicity windows and for the results of the ap-

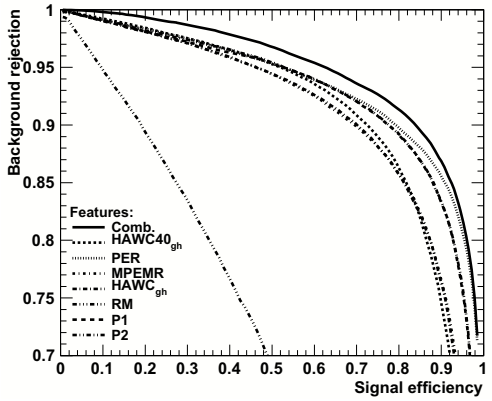

Figure 1: The signal efficiency dependent on the background rejection factor for the hit multiplicity window 100–200. The meaning of lines as shown in the inset.

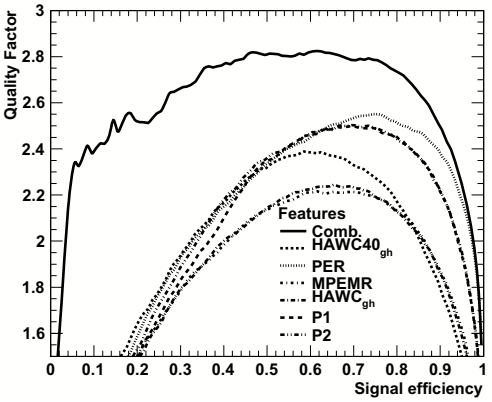

Figure 2: The dependence of parameter Qf on the signal efficiency for the hit multiplicity window 100–200.

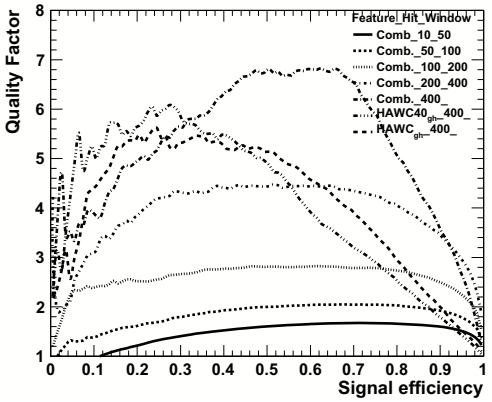

Figure 3: The dependence of parameter Qf on the signal efficiency for the all hit multiplicity window. The meaning of the lines is shown in the inset.

plication of all features together. We also show the results for the feature $HAWC40_{\rm gh}$ and $HAWC_{\rm gh}$ for comparison. As can be seen from the figure, combining multiple features significantly improves Qf for the hit multiplicity window > 400 in the signal high efficiency region. This improvement is almost a factor of two at the efficiency 75%. The $HAWC_{\rm gh}$ has slightly better performance than the $HAWC40_{\rm gh}$ and this is due to the usage of bigger R value for this hit multiplicity window.

3 Conclusions

This study has demonstrated that the feature already implemented in official HAWC software is a good starting point, which can still be optimized slightly. It is also shown that other features provide better performance in some hit multiplicity windows. The sensitivity of the detector can be enlarged by a factor of 1.2 on average when multiple features are combined. Large improvement is observed for the hit multiplicity window > 400.

Further studies are needed to define the optimal combination of features multiplicity window of the studied and to identify other features to improve the Qf parameter.

Acknowledgments V.G. is grateful to Yerevan physics institute to spent a sabbatical year in YerPhi. We acknowledge PASPA and PAPIIT IN-115409, IN-121309 DGAPA-UNAM for a partial financial support. This work has also been supported by the National Science Foundation, the U.S. Department of Energy, and CONACyT (Mexico).

References

[1] A.A. Abdo et al., ApJ **658**, (2007) L33–L36
[2] G. Aielli et al. (ARGO-YBJ Coll.), NIM **A 562**, (2006) 92–96
[3] M. DATTOLI et al., ARGO-YJB Collaboration, ICRC 2007 Merida, Mexico.
[4] Andrew J. Smith, for the HAWC Collaboration, PROCEEDINGS OF THE 31st ICRC, Łoódź 2009.
[5] D.Heck et al., Report FZKA 6019 Forschungszentrum Karlsruhe, (1998)
[6] S. Agostinelli et al, NIM **A 506** (2003) 250–303
[7] arxive 0811.1510 astro:ph
[8] http://tmva.sourceforge.net, arXiv:physics/0703039, CERN-OPEN-2007-007
[9] R. Brun and F. Rademakers, ROOT; NIM **A 389**, 81 (1997).
[10] Chilingarian A. A., Analysis and Nonparametric Inference http://crdlx5.yerphi.am/proj/ani.