Gamma-Hadron-Separation in the MAGIC Experiment

Tobias Voigt, Roland Fried, Michael Backes, and Wolfgang Rhode

Abstract The MAGIC-telescopes on the canary island of La Palma are two of the largest Cherenkov telescopes in the world, operating in stereoscopic mode since 2009 (Aleksić et al., Astropart. Phys. 35:435–448, 2012). A major step in the analysis of MAGIC data is the classification of observations into a gamma-ray signal and hadronic background. In this contribution we introduce the data provided by the MAGIC telescopes, which has some distinctive features. These features include high class imbalance, unknown and unequal misclassification costs as well as the absence of reliably labeled training data. We introduce a method to deal with some of these features. The method is based on a thresholding approach (Sheng and Ling 2006) and aims at minimization of the mean square error of an estimator, which is derived from the classification. The method is designed to fit into the special requirements of the MAGIC data.

1 Introduction

Binary classification problems are quite common in scientific research. In very high energy (VHE) gamma-ray astronomy for example, the interest is in separating the gamma-ray signal from a hadronic background. The separation has to be done as exactly as possible since the number of gamma-ray events detected is needed for the calculation of energy spectra and light curves (Mazin 2007). There are some distinctive features characterizing the data we have to deal with.

T. Voigt (✉) · R. Fried
Faculty of Statistics, TU Dortmund University, Vogelpothsweg 87, 44227 Dortmund, Germany
e-mail: voigt@statistik.tu-dortmund.de; fried@statistik.tu-dortmund.de

M. Backes · W. Rhode
Physics Faculty, TU Dortmund University, Otto-Hahn-Straße 4, 44227 Dortmund, Germany
e-mail: backes@physik.tu-dortmund.de; rhode@physik.tu-dortmund.de

M. Spiliopoulou et al. (eds.), *Data Analysis, Machine Learning and Knowledge Discovery*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-319-01595-8_13,

One feature is that there is a huge class imbalance in the data. It is known that hadron observations (negatives) are more than 100–1,000 times more common than gamma events (positives) (Weekes 2003; or Hinton and Hofman 2009). The exact ratio, however, is unknown. A second feature is that individual misclassification costs of gamma and hadron observations are unknown and not important in our context. We use classification as a preliminary step of an analysis, which aims at estimation of some quantity. The mean square error of the resulting estimator thus measures naturally also the expected loss resulting from our classification.

Throughout this paper we use random forests (Breiman 2001) as is usually done in the MAGIC experiment (Albert et al. 2008). One effective method of making these cost sensitive is the thresholding method (Sheng and Ling 2006). This method is not applicable as we do not know individual misclassification costs, but in the following we introduce a similar method based on the third feature of the data: In VHE gamma-ray astronomy one is not primarily interested in the best possible classification of any single event, but instead one wants to know the total number of gamma observations (positives) as this is the starting point for astrophysical interpretations. Statistically speaking this means estimation of the true number of positives based on a training sample. As said above the mean square error of this estimation measures naturally the expected loss of the classification, so we regard the mean square error (MSE) as overall misclassification risk in the thresholding method and choose the discrimination threshold which minimizes the MSE of the estimated number of positives in a data set. Additionally, the unknown class imbalance is taken into consideration by this method.

2 The MAGIC Experiment and Data

The MAGIC telescopes on the canary island of La Palma are two of the biggest Cherenkov telescopes in the world. Their purpose is to detect highly energetic gamma particles emitted by various astrophysical sources like Active Galactic Nuclei (AGNs). Gamma particles are of special interest to astrophysicists, because they are not scattered by magnetic fields, so that their point of origin can be reconstructed. When a gamma particle reaches Earth, it interferes with the atmosphere, inducing a so called air shower of secondary particles. The air shower then emits Cherenkov light in a cone, which can be seen by Cherenkov telescopes like the MAGIC telescopes. The somewhat elliptical shape of the shower is imaged in the telescopes' cameras. A major issue one has to solve in the MAGIC experiment is that not only gammas induce particle showers, but also many other particles, summarized as hadrons. Thus, gamma and hadron particles have to be separated through classification. Figure 1 shows camera images of the MAGIC I telescope of a gamma and hadron event. As can be seen, the gamma event has a more regular shape than the hadron. Note though that these images are almost ideal cases. Usually the difference between the two types of particles cannot be seen that easily. Figure 1 also shows the raw data we have for the analysis. It consists of one light intensity

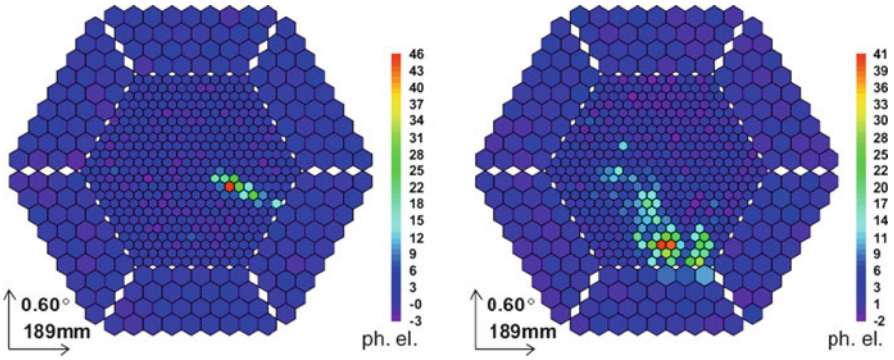

Fig. 1 Camera images of a gamma event (*left*) and a hadron event (*right*) in the MAGIC experiment

for each pixel in the camera. Additionally, but not shown here, a time information is given for each pixel.

One of the major goals of the MAGIC experiment is the Unfolding of Energy Spectra. Energy Spectra are basically histograms of the energy of observed gamma particles, that is an estimate of the unknown energy distribution of a source. From this distribution, characteristics of the source can be inferred. That means, what we are aiming for is to estimate the number of gamma observations in each of the histogram's energy bins as precisely as possible, to get a good estimation of the true energy distribution.

In order to achieve this goal one has to deal with some challenges.

Very Unfortunate Signal-Background Ratio

According to Weekes (2003) hadron events are around 100–1,000 times more common than gamma events. This leads to a very undesirable signal-to-background ratio for our classification. This high class imbalance makes classification of gammas and hadrons much more difficult than it could be with a more desirable ratio. Additionally, the true ratio is different for each source one is taking data from. So we cannot use any a priori knowledge about the ratio to make classification easier.

No Reliably Labeled Training Data

Another challenge one is facing in the analysis of MAGIC data is that we do not have access to training data to train the random forest. That means we cannot draw a sample from the joint distribution of gammas and hadrons with known labels. What we can do is to take real hadronic background data and mix it with simulated

gamma events to get training data. The difference of this to drawing from the joint distribution is that we cannot estimate the true gamma-hadron-ratio from the mix, as the number of gamma and hadron events in the mix is chosen manually. To estimate the number of gamma events in real data, it is however necessary to be able to assess this ratio. So we have to find a way to accomplish this.

Misclassification Costs

The third challenge is that we know that a misclassification of observations causes a worse estimation of the number of gamma events. That means, there are some misclassification costs, so that it is desirable to have a cost-sensitive classifier. A random forest, which we use in the MAGIC analysis chain, can be made cost sensitive in various ways. One is the thresholding method by Sheng and Ling (2006). The idea of this method is to minimize the misclassification costs over the classification threshold in the random forest's output, that is the fraction of votes of the trees. However, to apply this method it is of course necessary to know the misclassification costs. So to make the classifier cost sensitive, we must first assess the misclassification costs.

3 Threshold Optimization

An example of how we try to achieve a good estimation of the energy spectrum is the optimization of the threshold in the outcome of the random forest.

Problem Setup

The problem we are facing is a binary classification problem. We have a random vector of input variables $\mathbf{X} = (X_1, \ldots, X_m)^T$ and a binary classification variable Y. $\mathbf{X}$ and Y have the joint distribution $P(\mathbf{X}, Y)$. We neither know this distribution, nor can we make any justifiable assumptions about it. Additionally, in our application it is not possible to draw a sample from this distribution. We are, however, able to draw samples from P($\mathbf{X}$) as well as the conditional distributions $P(\mathbf{X}|Y = 0)$ and $P(\mathbf{X}|Y = 1)$. Thus, we have independent realizations $(\mathbf{x}_1, 0), \ldots, (\mathbf{x}_{n_0\cdot}, 0)$ and $(\mathbf{x}_{n_0\cdot+1}, 1), \ldots, (\mathbf{x}_{n_1\cdot+n_0\cdot}, 1)$ from the respective distributions with sizes $n_{0\cdot}$ and $n_{1\cdot}$, respectively, and $n = n_{0\cdot} + n_{1\cdot}$.

Many classifiers can be interpreted as a function $f : \mathbb{R}^m \to [0, 1]$. In the MAGIC experiment we use random forests, but any classifier which can be regarded as such a function f can be used. For a final classification into 0 and 1 we need a threshold c, so that

$$g(\mathbf{x}; c) = \begin{cases} 0, & \text{if } f(\mathbf{x}) \leq c \\ 1, & \text{if } f(\mathbf{x}) > c \end{cases}.$$

Table 1 True and classified numbers of positives and negatives in a training sample (left), in a sample of actual data (middle) and in Off data (right)

		Classified		
		1	0	$\sum$
True	1	n_{11}	n_{10}	$n_{1\cdot}$
	0	n_{01}	n_{00}	$n_{0\cdot}$
	$\sum$	$n_{\cdot 1}$	$n_{\cdot 0}$	n

		Classified		
		1	0	$\sum$
True	1	N_{11}	N_{10}	$N_{1\cdot}$
	0	N_{01}	N_{00}	$N_{0\cdot}$
	$\sum$	$N_{\cdot 1}$	$N_{\cdot 0}$	N

		Classified		
		1	0	$\sum$
True	1	0	0	0
	0	N_{01}^{off}	N_{00}^{off}	$N_{0\cdot}^{off}$
	$\sum$	$N_{\cdot 1}^{off}$	$N_{\cdot 0}^{off}$	N^{off}

We consider f to be given and only vary c in this paper. There are several reasons why we consider f to be given. Among other reasons, we want to adapt the thresholding method by Sheng and Ling (2006) and we cannot change the MAGIC analysis chain too drastically, as all changes need approval of the MAGIC collaboration.

In addition to the training data, we have a sample of actual data to be classified, $\mathbf{x}_1^*, \ldots, \mathbf{x}_N^*$, for which the binary label is unknown. This data consists of N events with $N_{1\cdot}$ and $N_{0\cdot}$ defined analogously to $n_{1\cdot}$ and $n_{0\cdot}$, but unknown.

As we have stated above, we only have simulated gamma events as training data and therefore need additional information to assess the gamma-hadron-ratio in the real data. This additional information is given by Off data, which only consists of $N_{0\cdot}^{off}$ hadron events. $N_{0\cdot}^{off}$ can be assumed to have the same distribution as $N_{0\cdot}$, so the two sample sizes should be close to each other. In fact, a realization of $N_{0\cdot}^{off}$ is an unbiased estimate for $N_{0\cdot}$. From this Off data we are able to estimate the true gamma-hadron-ratio.

For a given threshold c we denote the numbers of observations in the training data after the classification as $n_{ij}, i, j \in \{1, 0\}$, where the first index indicates the true class and the second index the class as which the event was classified. We can display these numbers in a 2×2 table. With $N_{ij}, i, j \in \{1, 0\}$ and $N_{0j}, j \in \{1, 0\}$ defined analogously we get similar 2×2 tables for the actual data and the Off data, see Table 1.

It is obvious that we do not know the numbers in the first two rows of the table for the actual data as we do not know the true numbers of positives and negatives $N_{1\cdot}$ and $N_{0\cdot}$.

As we can see above, $N_{1\cdot}$ is considered to be a random variable and our goal is to estimate, or perhaps better predict, the unknown realization of $N_{1\cdot}$. The same applies to $N_{0\cdot}$. That is why we consider all the following distributions to be conditional on these values.

We additionally define the True Positive Rate (*TPR*), which is also known as Recall or (signal) Efficiency, and the False Positive Rate (*FPR*) as

$$TPR = \frac{n_{11}}{n_{1\cdot}} \tag{1}$$

and

$$FPR = \frac{n_{01}}{n_{0\cdot}}, \tag{2}$$

respectively. As we will see in the following section, these two values are important in the estimation of the number of gamma events.

Estimating the Number of Gamma-Events

To estimate the number of gamma events, we first have a look at the following estimator:

$$\tilde{N}_{1\cdot} = \frac{1}{p_{11}} \left(N_{\cdot 1} - N_{01}^{off} \right). \tag{3}$$

where p_{11} is the (unknown) probability of classifying a gamma correctly. This estimator could be used if we knew p_{11}. It takes the difference between $N_{\cdot 1}$ and N_{01}^{off} as an estimate for N_{11} and multiplies this with $\frac{1}{p_{11}}$ to compensate for the classification error in the signal events.

Since we want to estimate the number of positives as precisely as possible we want to assess the quality of the estimator $\tilde{N}_{1\cdot}$. A standard measure of the quality of an estimator is the mean square error (MSE). As in applications we usually have fixed samples in which we want to estimate $N_{1\cdot}$, we calculate the MSE conditionally on $N_{1\cdot}$, $N_{0\cdot}$ and $N_{0\cdot}^{off}$. Under the assumption that N_{i1}, N_{01}^{off} and n_{i1}, $i \in \{1, 0\}$ are independent and (conditionally) follow binomial distributions, the conditional MSE of $\tilde{N}_{1\cdot}$ can easily be calculated. It is:

$$\begin{aligned} \mathrm{MSE}\left(\tilde{N}_{1\cdot} | N_{1\cdot}, N_{0\cdot}, N_{0\cdot}^{off} \right) = {} & \frac{p_{01}^2}{p_{11}^2} \left(N_{0\cdot} - N_{0\cdot}^{off} \right)^2 \\ & + N_{1\cdot} \left(\frac{1}{p_{11}} - 1 \right) + \frac{p_{01} - p_{01}^2}{p_{11}^2} \left(N_{0\cdot} + N_{0\cdot}^{off} \right) \end{aligned} \tag{4}$$

where p_{01} is the probability of classifying a hadron as gamma and p_{11} is the probability of classifying a gamma correctly. As we do not know these values we have to estimate them. Consistent estimators for these values are *TPR* and *FPR* [(1) and (2)]. Using *TPR* as an estimator for p_{11} in (3) we get

$$\hat{N}_{1\cdot} = \frac{n_{1\cdot}}{n_{11}} \left(N_{\cdot 1} - N_{01}^{off} \right) = \frac{1}{TPR} \left(N_{\cdot 1} - N_{01}^{off} \right). \tag{5}$$

By estimating p_{11} with *TPR* and p_{01} with *FPR* in (4) we get the estimate

$$\begin{aligned} \widehat{\mathrm{MSE}}\left(\hat{N}_{1\cdot} | N_{1\cdot}, N_{0\cdot}, N_{0\cdot}^{off} \right) = {} & \frac{FPR^2}{TPR^2} \left(N_{0\cdot} - N_{0\cdot}^{off} \right)^2 \\ & + N_{1\cdot} \left(\frac{1}{TPR} - 1 \right) + \frac{FPR - FPR^2}{TPR^2} \left(N_{0\cdot} + N_{0\cdot}^{off} \right). \end{aligned} \tag{6}$$

As *TPR* and *FPR* are consistent estimators of p_{11} and p_{01} and the sample sizes $n_{1\cdot}$ and $n_{0\cdot}$ are usually high ($> 10^5$), using the estimates instead of the true probabilities should only lead to a marginal difference.

Algorithm

Equations (5) and (6) can be used in an iterative manner to find a discrimination threshold, although $N_{1\cdot}$ in (6) is unknown. To find a threshold we alternately estimate $N_{1\cdot}$ and calculate the threshold:

1. Set an initial value c for the threshold.
2. With this threshold estimate $N_{1\cdot}$ using equation (5).
3. Compute a new threshold through minimizing equation (6) over all thresholds using the estimates $\hat{N}_{1\cdot}$ for $N_{1\cdot}$ and $N - \hat{N}_{1\cdot}$ for $N_{0\cdot}$.
4. If a stopping criterion is fulfilled, compute a final estimate of $N_{1\cdot}$ and stop. Otherwise go back to step 2.

Because negative estimates $\hat{N}_{1\cdot}$ can lead to a negative estimate of the MSE, we set negative estimates to 0. As a stopping criterion, we require that the change in the cut from one iteration to the next is below 10^{-6}. First experiences with the algorithm show that the convergence is quite fast. The stopping criterion is usually reached in less than ten iterations.

In the following we refer to this algorithm as the MSEmin method. This method takes both into consideration: The problem of class imbalance and the minimization of the MSE, that is, the overall misclassification costs. In the next section we investigate the performance of this algorithm on simulated data and compare it to other possible approaches.

4 Application

It is now of interest, if the MSEmin method proposed above means an improvement over the currently used and other methods. As stated above, we want to estimate the number of gamma events depending on the energy range. We therefore use the methods on each of several energy bins individually by splitting the data sets according to the (estimated) energy of the observations and using the methods on each of these subsamples individually.

The method currently in use in the MAGIC experiment is to choose the threshold manually so that the *TPR* is "high, but not too high". Often *TPR* is set to values between 0.4 and 0.9 (e.g. Aleksić et al. 2010) and the threshold is chosen accordingly. For our comparison we look at values of 0.1–0.9 for *TPR*. We call these methods Recall01, ..., Recall09.

An approach to avoid energy binning is to fit a binary regression model to the random forest output, with energy as the covariate. The fitted curve can then be

regarded as discrimination threshold. In this paper we use a logistic regression model. We fit the model to the training data using a standard Maximum Likelihood approach to estimate the model coefficients.

As the proposed MSEmin method is quite general and not bound to optimizing a fixed threshold, we use it in an additional approach by combining it with logistic regression. Instead of minimizing the MSE over possible fixed thresholds, we search for optimal parameters of the logistic regression curve, so that the MSE becomes minimal. The procedure is the same as for the MSEmin method proposed in the algorithm above, only that we exchange the threshold c with the two parameters of the logistic regression, say β_0 and β_1. For initialization we use the ordinary ML-estimates of the two parameters.

We use all these methods on 500 test samples and check which method gives the best estimate for the number of gamma events. We focus here on the hardest classification task with a gamma-hadron-ratio of 1:1000. For the comparison we use the following data:

Test data: To represent actual data we simulate 500 samples for each gamma-hadron-ratio 1:100, 1:200, ..., 1:1000. The number of hadron-events in each sample is drawn from a Poisson distribution with mean 150,000. The number of gamma events is chosen to match the respective ratio.

Training data: We use 652,785 simulated gamma observations and 58,310 hadron observations to represent the training data from which *TPR* and *FPR* are calculated. Note that the ratio of gammas and hadrons in this data has no influence on the outcome, as only *TPR* and *FPR* are calculated from this data.

Off data: For each test sample we draw a sample of hadron observations to represent the Off data. The number of hadrons in each sample is drawn from a Poisson distribution with mean 150,000.

The result can be seen in Fig. 2. As we can see, all methods seem to give unbiased estimates. However, all Recall methods have comparably high variances when estimating the true number of gamma events, with the best one being Recall01. Our proposed method MSEmin leads to a smaller variance and therefore performs better than all of them. The results of the logistic regression approach is quite similar to the MSEmin method, but has a bit smaller errors. The best performance is given by the combination of the methods MSEmin and logistic regression.

5 Conclusions and Outlook

MAGIC data has some distinctive features, which make the analysis of the data difficult. We have illustrated that major challenges can be overcome when we focus on the overall aim of the analysis, which is the estimation of the number of signal events.

We introduced a method to choose an optimal classification threshold in the outcome of a classifier, which can be regarded as a function mapping to the interval

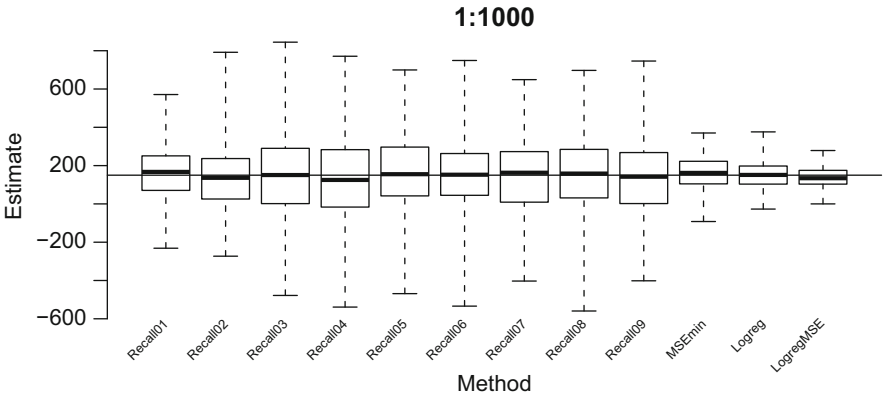

Fig. 2 Boxplots of the estimates in the 500 samples with a gamma-hadron-ratio of 1:1000. The *thick line* in the middle of each box represents the median of the estimates. Between the *upper* and *lower boundaries* of each box lie 50 % of the estimates. The whiskers range to the minimum and maximum of all the data. The true number is marked by the long *horizontal line*

[0,1]. In this paper we used random forests, but any classifier providing such a function can be used. The introduced method minimizes the MSE of the estimation of the number of signal events. In our experiments this method performs better than the method currently used. The method is also adaptable to combine it with other methods. The combination with a logistic regression approach gave even better results than the two methods on their own.

Acknowledgements This work has been supported by the DFG, Collaborative Research Center SFB 876. We thank the ITMC at TU Dortmund University for providing computer resources on LiDo.

References

Albert, J., et al. (2008). Implementation of the random forest method for the imaging atmospheric Cherenkov telescope MAGIC. *Nuclear Instruments and Methods in Physics Research A*, *588*, 424–432.

Aleksić, J., et al. (2010). MAGIC TeV gamma-ray observations of Markarian 421 during multiwavelength campaigns in 2006. *Astronomy and Astrophysics*, *519*, A32.

Aleksić, J., et al. (2012). Performance of the MAGIC stereo system obtained with Crab Nebula data. *Astroparticle Physics*, *35*, 435–448.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5.

Hillas, A. M. (1985). Cherenkov light images of EAS produced by primary gamma. In *Proceedings of the 19th International Cosmic Ray Conference ICRC* (Vol. 3, p. 445), San Diego.

Hinton, J. A., & Hofman, W. (2009). Teraelectronvolt astronomy. *Annual Review of Astronomy & Astrophysics*, *47*, 523–565.

Mazin, D. (2007). *A Study of Very High Energy-Ray Emission From AGNs and Constraints on the Extragalactic Background Light*, Ph.D. Thesis, Technische Universitaet Muenchen.

Sheng, V., & Ling, C. (2006). Thresholding for making classifiers cost sensitive. In *Proceedings of the 21st National Conference on Artificial Intelligence* (Vol. 1, pp. 476–481). AAAI Press, Boston.

Thom, M. (2009). *Anaylse der Quelle 1ES 1959 + 650 mit MAGIC und die Implementierung eines Webinterfaces fr die automatische Monte-Carlo-Produktion*, Diploma Thesis, Technische Universitaet Dortmund.

Weekes, T. (2003). *Very High Energy Gamma-Ray Astronomy*. Bristol/Philadelphia: Institute of Physics Publishing.