MULTIDIMENSIONAL EVENT CLASSIFICATION IN IMAGES FROM GAMMA-RAY AIR SHOWERS

R.K.Bock, W.Wittek
MPI Munich

Abstract

Exploring signals from outer space has become an observational science under fast expansion: astroparticle physics. Among earthbound observations, the technique of imaging gamma-ray Cherenkov telescopes using the atmosphere as calorimeter is a particularly recent technique. Events in such telescopes appear as 2-dimensional images, and the image characteristics have to be used to discriminate between the interesting gammas and the dominating charged particles, mostly protons. Present techniques of analysis express the images in terms of several parameters; the goal is to find some test statistic(s) which allow(s) to optimize the classification. Among the available classification techniques, several have been used for a comparative study, and more are under investigation. The methods will be briefly presented, remaining problems are discussed, some preliminary results are shown.

1 Cherenkov telescopes - one of many tools in astrophysics

Astronomy and astrophysics have made rapid progress as an observational science in recent years. They capture today particles of different nature and of very different energy, more or less whatever the galaxy and the universe spray on us - within the limits set by instrumentation. Beyond the venerable astronomical observations at visible wavelengths, with a centuries-long history, and those of (charged) Cosmic Rays, started less than a hundred years ago, one observes now the electromagnetic spectrum over some 20 orders of magnitude, from radio over microwave and optical wavelengths to X-rays and the highest energy gamma-rays (see figures 1 and 2). In the very recent past, major neutrino detectors have been built or are under construction that will also contribute important information, or have already done so.

For the electromagnetic spectrum, the absorption in the earth's atmosphere varies substantially in function of wavelength; for most wavelengths outside the optical and radio bands, instruments have to be flown on satellites. This was also where the first gamma observations were made. Ground-based atmospheric Cherenkov telescopes using the imaging technique are a comparatively recent addition to the panoply of successful instruments, with the first results demonstrated not before 1985. They can be built with a much larger effective collection area than detectors sent into orbit, and hence respond to the lower fluxes of higher-energy primary gammas. On the other hand, the number of observable Cherenkov photons for lower energy (below 100 GeV) primary photons becomes comparatively small, and correspondingly the problems of discrimination against background get enhanced; this translates for the detector into general requirements of optimal light collection and of best possible image processing. This note deals with one aspect of discrimination. For an early review see [1].

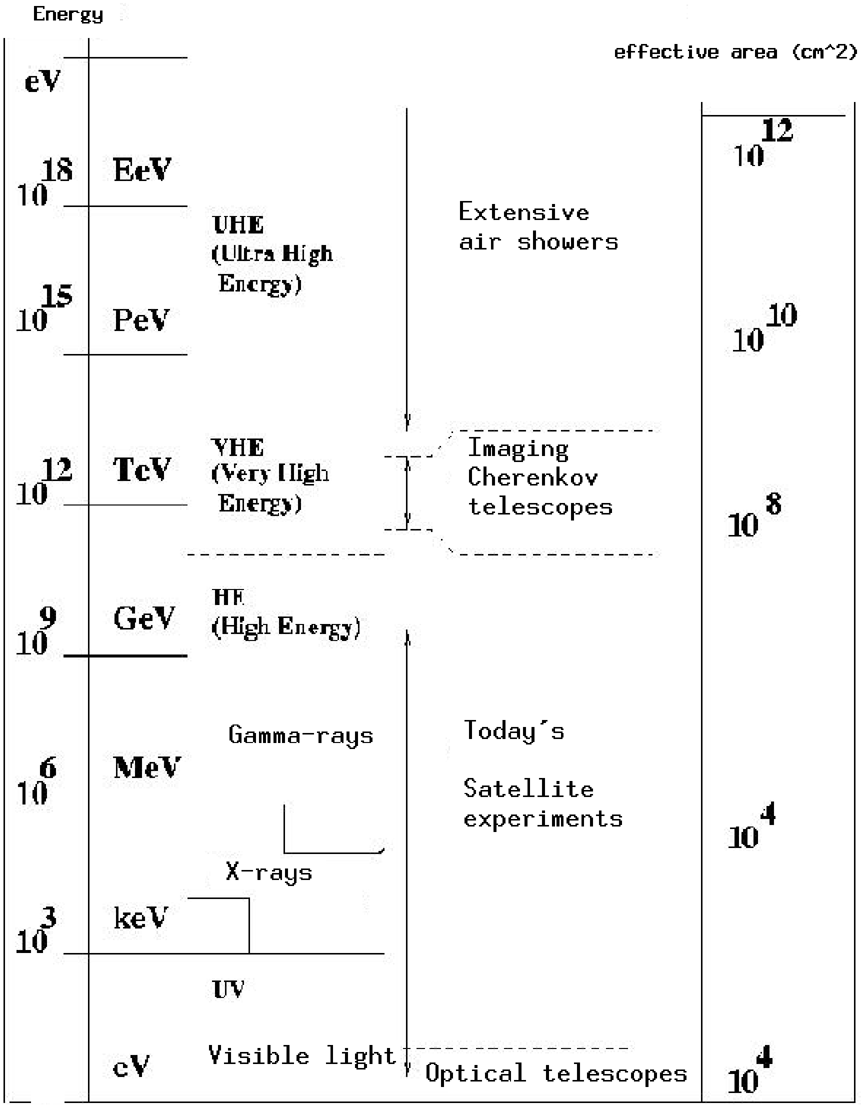

Fig. 1: The observational methods for the electromagnetic sky

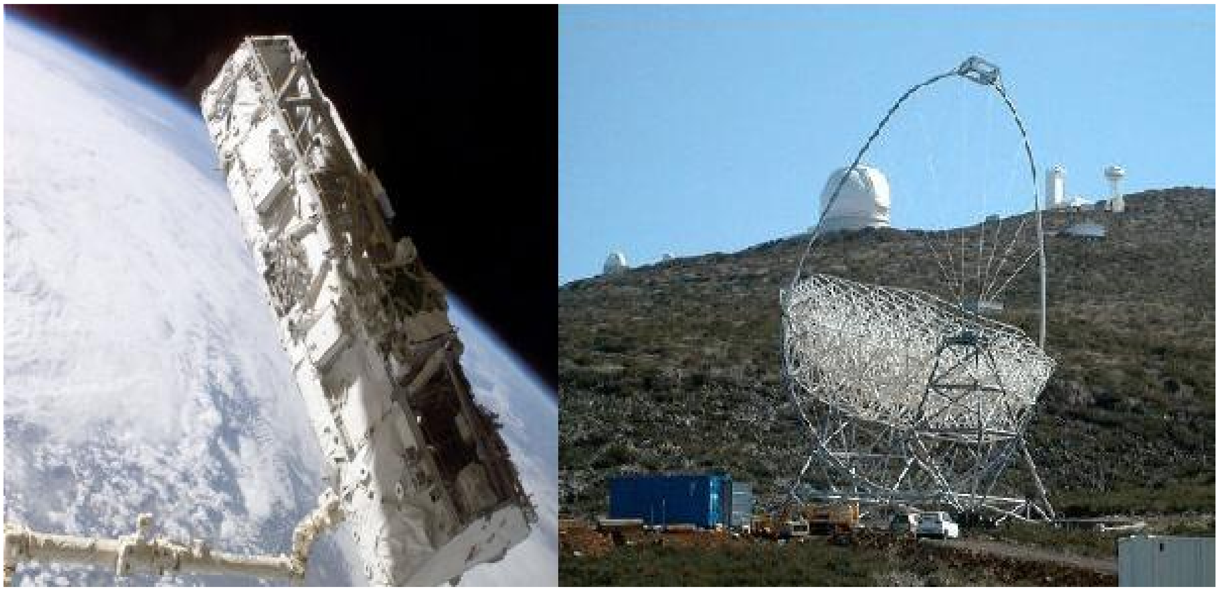

Fig. 2: Satellite-based instruments (left), and the future largest imaging gamma-ray telscope, MAGIC (right)

2 Principle of imaging Cherenkov telescopes

Cherenkov telescopes take advantage of the radiation emitted by charged particles as they are produced abundantly inside showers developing in the atmosphere. The showers for incoming photons produce rather exclusively electrons, which show strong Cherenkov radiation, and more photons. The showers absorb the initial photon like any electromagnetic calorimeter would, but the leakage radiation allows reconstruction of its most relevant parameters. As it happens, it is not only the interesting primary photons that cause showers in the atmosphere: also the ubiquitous cosmic rays, mostly protons, tend to shower and produce relativistic particles which can mimic a gamma- initiated shower. Cherenkov radiation is emitted in the visible to UV regime, so a Cherenkov telescope uses visible light, but has to make best use of relatively few light quanta that arrive on the ground and can be converted to a signal. The term ’imaging’ takes here the meaning of localizing the incoming photons over the very short period the shower front sweeps over the telescope (1-2 nsec). Depending on the energy of the primary photon, some 100 to 1000 Cherenkov photons get collected. The main analysis problem consists of discriminating the image characteristics found for primary photons from those of debris left behind by a hadronic shower in the atmosphere. High sensitivity and good time resolution are vital properties for an imaging gamma-ray telescope; spatial precision, on the other hand, is not mandatory because the observed photons have a natural spread, due to the showering process and the limits imposed by the atmosphere (this is not a constant calorimeter!). Consequently, high reflectivity mirrors are important, and so are the best possible photomultipliers in the camera, but the precision of focus (unlike in an optical telescope) need not be much better than several millimeters in the plane of photomultipliers (or of the order of one minute of arc in angle).

3 Principle of image parameters

As the sources of high-energy gammas are few and their signal comparatively weak, it is the Cherenkov photons from hadron showers that dominate the hardware trigger, and some involved image analysis must achieve the discrimination between primary gammas and hadrons. Fortunately, these showers show different characteristics (like in any calorimeter); with suitable feature extraction the pixels making up the image can be converted to some set of image parameters which statistically allow a separation of events. Historically, one uses some image cleaning algorithm to eliminate outliers, viz. retain only pixels touched by Cherenkov photons from the showering, and then executes a principal component analysis (see [2]) in the camera plane. That results in a correlation axis and defines an ellipse (see fig. 3) - if the deposition were distributed as a bivariate Gaussian, this would be the error ellipse; in reality, the images of primary photons indeed tend to resemble an ellipse, albeit systematically asymmetric along its major axis. The characteristic parameterss of this ellipse (often called Hillas parameters, [1]), plus several ad-hoc parameters describing asymmetries and deviations from a smooth spectrum constitute the image characteristics to be used.

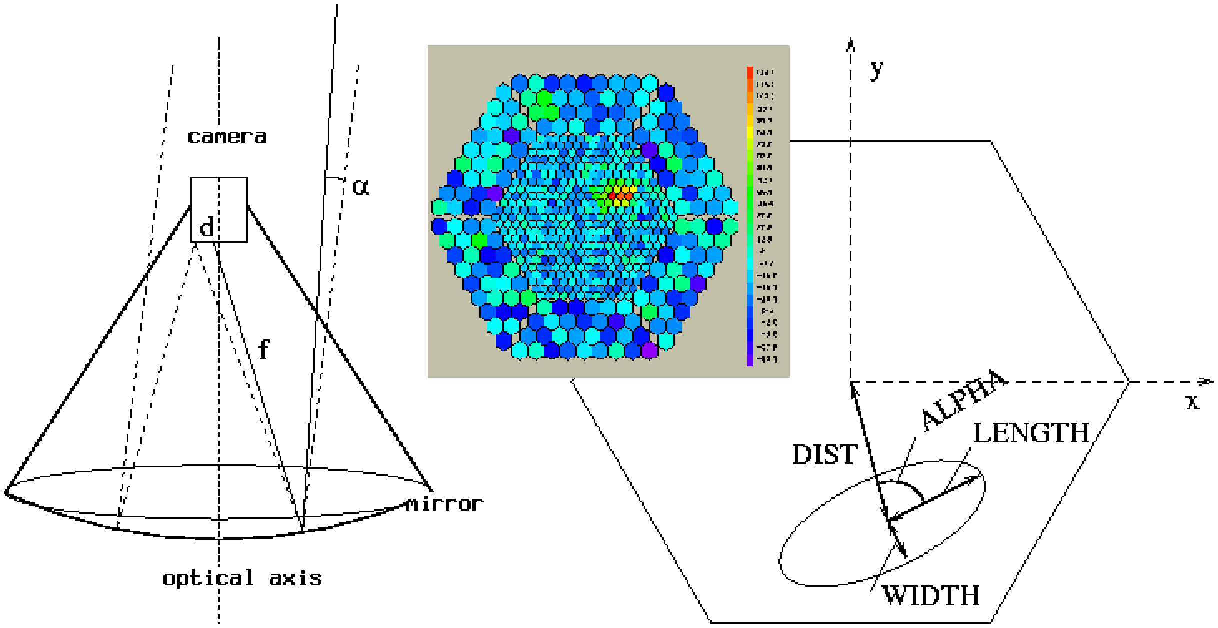

Fig. 3: Cherenkov telescope: sketch and image parameters, and a simulated event in MAGIC (inset)

The optimization of the parameters requires detector-dependent studies; the optimization of the discrimination to be applied to a given set of parameters, on the other hand, can be studied in more generality, and is the goal of an ongoing exercise described in this contribution. For this, we use a data set generated by a Monte Carlo program. Data are computed for the future MAGIC telescope, which is located on the Canary island of La Palma, and will become operational in 2002.

4 Multivariate classification

If confronted with a single test statistic, showing a one-dimensional probability density different for signal and background events (in our case gammas and hadrons, respectively), the discrimination is simple: by applying a cut one selects all events with this parameter larger (or smaller) than the cut value; for a small enough cut value one obtains zero acceptance for both signal and background, for a large enough cut value an acceptance equal to one for both samples; cut values chosen in between will make the two acceptances lie on a line deviating from equal acceptance for the two samples the more the better the variable was chosen, i.e. acceptances are higher for signal than for background events (see fig.4). The diagram of signal vs. background acceptance is known under the name Neyman-Pearson diagram or decision quality diagram. Note that these acceptances are directly related to statistical quantities like purity, cost, contamination etc.

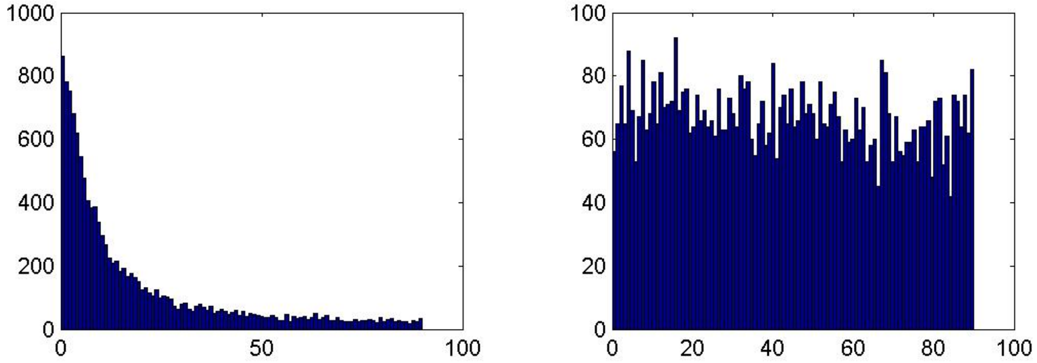

Fig. 4: Example of a good discrimination variable (Alpha). Left: signal, right: background

The problem becomes more involved when faced with multiple variables, the *multivariate discrimination* problem. Different classification methods for the multivariate case are in use, but little guidance exists as to their advantages or pitfalls; it is our objective to apply some of them to fixed samples of test data and define rigid criteria for comparing the results.

5 Different classification methods

General multivariate classification methods are advertised in puzzling numbers, some of them are available on the commercial market; we have started to compare several methods below; others which we have not (yet?) tried out are variants of discriminant analysis (QDA = quadratic d.a., RDA = regularized d.a., DASCO = discriminant analysis with shrunken covariances), SVM (a learning system called support vector machines), or variants of the kernel method, like adaptive kernels. Also, we show for the moment no results on ANN (artificial neural network) methods.

5.1 Cuts and supercuts

Cuts can be applied in the n-space of features (in our case image parameters), one variable at a time or logically related with AND or OR; the problem gets unwieldy even at low n, nevertheless, this is the most commonly used method amongst physicists. Wide experience exists also for all operating Cherenkov telescopes (e.g. [1], [3]); any method claiming to be superior must use results from these as a yardstick. Working in 1-dimensional projections, correlations, in particular non-linear, between variables are not easy to take into account, although cuts in a variable are sometimes made dependent on the value of other parameters (*supercuts* or *dynamic cuts*). Decorrelation by standard methods (Principal Component Analysis, [2]) does not solve the problem in general, being a linear operation. Defining new variables does help, so do the dynamic cuts. The cut method also does need an optimization criterion, and will not result in a relation between gamma acceptance and hadron acceptance, i.e. no single test statistic is defined. Usually, cut optimization leads to separate studies and approximations for each new data set (this is based on past experience), which makes results sometimes difficult to reproduce. Cutting needs very clearly independent samples for training and control.

5.2 classification and regression trees (CART)

This is a method or rather a family of methods developed originally by high-energy physicists to do away with the randomness in optimizing cuts ([4]); it has now developed into a data mining method, with tools available commercially from several vendors. The basic operations are alternate sequences of growing a decision tree and pruning it, done according to slowly adapting criteria in some heuristic succession until some convergence criterion is satisfied. CART methods need very clearly independent samples for training and control. The main problem is to find a robust measure to choose from the many trees that are (or can be) grown, avoiding overtraining (e.g. refining the tree until each event in the training sample has its own sequence of cuts). CART is made for large samples, there is no experience with Cherenkov telescope data, but some promising early results have been produced (see preliminary results in fig.5).

5.3 Linear discriminant analysis (LDA)

This is a popular method because it results in an elegant parametric calculation. Its objective is to find a linear combination of the original image parameters such that the hyperplane defined by the transformation maximizes the distance between the means of signal (gamma) and background (hadron) samples, simultaneously minimizing the variance inside each sample.The method is fast, simple and (probably) very robust. It does not depend on training samples. However, it ignores non-linear correlations in n-dimensional space (because of the linear transformation), and very little practical experience with LDA in Cherenkov telescope data exists; early tests show that at least higher-order variables are needed (e.g. x, y could provide the additional variable x^2y).

The formalism is simple ([5]): the transformation into the 'best separable space' is performed by the eigenvectors of a matrix readily derived from the data (for our application: in two classes, gammas g and protons p). Given samples $g_i(i = 1, n_g)$ for gammas and $p_j(j = 1, n_p)$ for protons, with nvar elements each, find a linear transformation vector a such that the transformed samples are $g' = a \cdot g$ and $p' = a \cdot p$, and the discriminating power

$$d = \frac{y^T S_b y}{y^T (S_b + S_w) y}$$

gets maximized, where y is the joint set of g' and p'. S_b(between-class variance) and S_w (within-class variance) are defined by:

$$S_w = \sum_{obs} (x_i - \mu_{class})(x_j - \mu_{class}), S_b = \sum_{class-1} (\mu_i - \mu_{tot})(\mu_j - \mu_{tot}),$$

where μ_{class} = class mean, μ_{tot} = overall mean, and x is the joint set of g and p. This leads, for two classes, to the result:

$$a = \frac{\sqrt{n_g n_p}}{(n_g + n_p)} (S_b + S_w)^{-1} (\mu_{tot_1} - \mu_{tot_2})$$

Like Principal Component Analysis (PCA, [2])), LDA is used for dimensionality reduction. LDA maximizes the ratio of between-class variance to within-class variance, for any pair of data sets. The goal of LDA is sample separability. PCA, on the other hand, minimizes the variances along new axes. The prime difference in application between LDA and PCA is that PCA performs feature classification (e.g. the image parameters from Cherenkov telescope data) while LDA performs sample classification. PCA changes both the shape and location of the data in its transformed space, whereas LDA provides more class separability by building a decision region between the classes. On our MAGIC test data set, LDA does not perform as well as other methods (see preliminary results in fig.5).

5.4 Kernel methods

The kernel density estimation is a nonparametric multivariate classification technique. The advantage is that of generality of the class-conditional and consistently estimated densities. Applied to our classification problem, the kernel function is used to evaluate an individual event likelihood, defined as the closeness to the population of gamma events or hadron events in n- dimensional space. The closeness is expressed by a kernel function applied to a reference sample. The method does not require a strict separation of training and control samples. The method has been toyed with in Whipple (the earliest functioning Cherenkov telescope), results look convincing ([6] and [7]); however, Whipple today still uses supercuts for analyzing its data; we have exposed our MAGIC test data to a basic kernel analysis, and the results look encouraging (see preliminary results in fig.5).

We define reference samples $g_i(i = 1, n_g)$ for gammas and $p_j(j = 1, n_p)$ for protons: Monte Carlo gammas, and 'off' events (obtained by measuring slightly off the source, to define the cosmics background in that region of the sky). We then find as classifier a likelihood function

$$R_g = k_g/k_p$$

with the kernel function $k_{g,p} = \sum_{g,p} k(x - x_r)$. x is the point (g or p) under consideration, x_r are the points in the reference sample (gamma or proton). The trick is to define a valid kernel function. Whipple has used a multivariate Gaussian (like a point spread function):

$$k \equiv \frac{exp(-(x - x_r)^T C_r^{-1} (x - x_r)/2)}{\sqrt{2}\pi^n \mid C_r \mid},$$

where C_r are the covariance matrices of the variables in the reference samples, n is the dimensionality.

The method needs comparing every event with every event in both reference samples, and thus is computationally costly. Whipple has reduced the parameter space and precomputed the kernel function for a lattice, using interpolation ([6] and [7]).

5.5 Composite probabilities

This home-developed method uses event probabilities obtained by comparing the event data to two- dimensional probability densities obtained from a training sample. Densities are determined by histogramming in two dimensions using bins that give constant bin content for signal data. All 2D projections are used that can be made from image parameters. Each bin thus has a probability to be signal, and a probability to be background; the probabilities of a given event for all projections are multiplied, and their product, the composite probability, is a single test statistic ('signalness'). The method was applied to some reference data from running experiments (Whipple, Hegra CT1), and (unpublished) results at least did match best existing results; strict comparisons suffered from moving data sets. The method needs some assumptions: number of bins, choice of training sample; our values were empirically chosen to be in an area where results are robust for the samples under study. Results are encouraging (see preliminary results in fig.5).

5.6 Artificial neural networks (ANN-s)

This method has been presented often in the past - it resembles the CART method but works in locally linearly transformed data. Usually, the results match but are not superior to whatever reference results existed. So far, no convincing case has been made for Cherenkov telescope data, although Whipple have tried the method (and remained with their supercuts). We have as yet no results for the MAGIC test data, but work is going on. There is substantial randomness in choosing the topology of the net, in particular the depth of the tree, the number of nodes, the training method, transfer function, etc.

6 Preliminary results

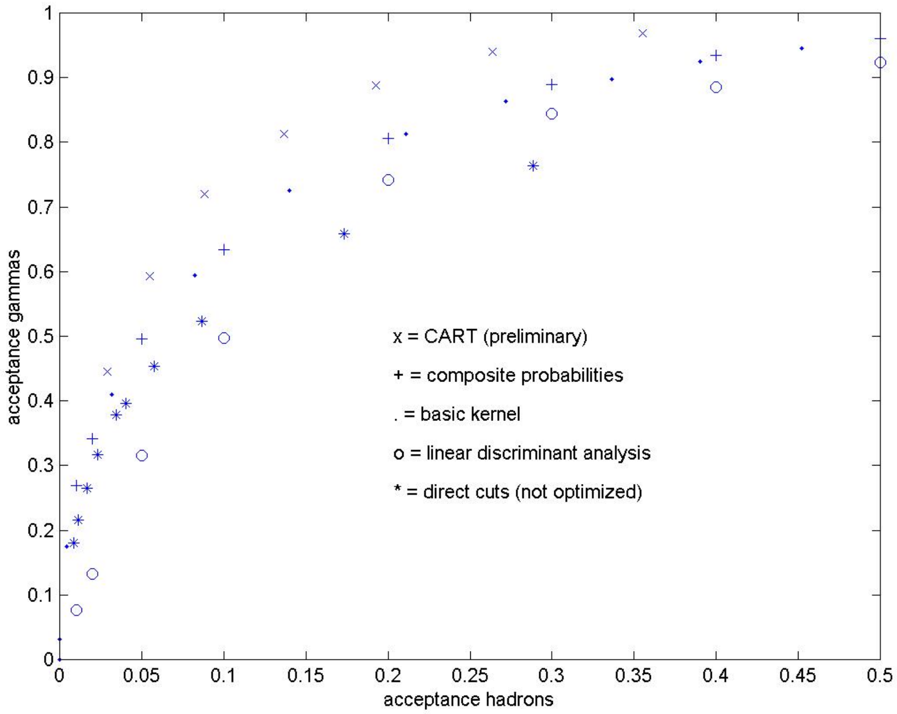

Fig. 5: Several classification methods compared

First results are shown in figure 5; the diagram shown relates acceptances for signal and background: the Neyman-Person diagram mentioned earlier; once the sample sizes are fixed, those two numbers can be translated into parameters that one usually wants to optimize, like statistical significances or other quality measures.

7 Criteria for a comparative study

In order to provide ourselves with an informed ’feeling’ for the advantages and pitfalls of various methods, we have defined a data set and rules for its analysis, that may be used by any of the methods. The rules include the definition of rigidly defined disjoint training and control samples, the requirement to give as a result estimators for corresponding hadron contamination and gamma acceptance (purity and cost), ideally a relation between the two i.e. a single test statistic; where this can not be done, results must be given for several optimization criteria, e.g. estimated hadron contamination at fixed gamma acceptance values, significance, etc. The advantage of working with Monte Carlo events is, of course, that results can be controlled by using the classification method on mixed samples and comparing the outcome to the known origin of events. We also attempt to add in our sample parameters which are not used for classification, but which can be used to show robustness. Data and criteria are available from the authors.

8 Possible conclusions and caveats

A systematic comparison of methods as we intend to perform, may give a conclusive result for the given data set. In all likelihood, some of the methods under study will show to be superior to direct cuts in parameters, used so far in all experiments. It is also likely that not a single one will turn out to be superior in all aspects, and we may conclude that the upper limit reached corresponds to the theoretically achievable limit. Eventually, among the best performers, ease of implementation, robustness, computer time and maybe other secondary criteria may have to decide, or we may choose to use several of them in parallel. Whatever conclusions are found will be valid for our input data, Monte Carlo events for the MAGIC telescope; the extrapolation of those conclusions to different data samples like real MAGIC data, different Cherenkov telescopes, or to a completely different problem must be validated anew. A real generalization to other problems can, obviously, not be addressed by a study like the intended one; its publication may, however, facilitate a future validation process, and allow us to discard some of the inferior methods.

The methods under study all assume an abstract space of image parameters, which is fine in Monte Carlo situations - and maybe only there: real data are subject to influences that distort this space. In our case the starfield in the field of view and the night sky background change during observation, the atmospheric conditions vary considerably, unavoidable detector changes and malfunction will occur, and all these effects may need corrections. We project, for our final analysis, to have gammas from Monte Carlo calculations, and measurements made on and off source: we must be able, therefore, to adapt the Monte Carlo variables for gammas to the prevailing observational conditions. For this problem of distortions of parameter space, some compromise between parametric corrections to the variables and frequent Monte Carlo computations for different observation conditions, is the likely solution.

No classification method can, of course, invent new independent parameters containing more information; they may be derived from the image, but could also be new independent observations, e.g. arrival time: to find these requires intuition in physics and good understanding of the detector.

References

[1] D.J.Fegan: Gamma/hadron separation at TeV energies, Topical Review, J.Phys.G: Nucl. Part. Phys. 23 (1997) 1013.

[2] I.T.Jollife: Principal Component Analysis, Springer New York, 1986. Many textbooks explain PCA, e.g. C.M.Bishop: Neural Networks for Pattern Recognition, Clarendon Press Oxford, 1995, or the web site http://www.statsoftinc.com/textbook/stathome.html

[3] D.Kranich: The temporal and spectral characteristics of the active galactic nucleus Mkn 501 during a phase of high activity in the TeV range, Dissertation an der Technischen Universität München, 2001.

[4] L.Breimann, J.H.Friedmann, R.A.Olshen, C.J.Stone: Classification and Regression Trees, Wadsworth, 1983.

[5] web site http://www.statsoftinc.com/textbook/stathome.html, look for discriminant function analysis. For an application in HEP: B.Fabbri, LDA with stepwise method, ALEPH 97-012 (internal report).

[6] P.Moriarty and F.W.Samuelson: Kernel Analysis in TeV Gamma-Ray Selection, Gamma-Ray Astrophysics Workshop, Snowbird, Utah 1999, AIP Conference Proceedings 515 p.338.

[7] S.Dunlea, P.Moriarty, and D.J.Fegan: Selection of TeV Gamma-Rays using the Kernel multivariate technique, Proceedings of the ICRC 2001, Copernicus Gesellschaft, Hamburg, p.2939.