Electron/pion separation in the CALICE WHCAL prototype using multi variate analysis techniques

Justas Zalieckas

August 2011

Abstract

This note describes the usage of TMVA 4 (Toolkit for Multivariate Data Analysis with Root) within a Marlin processor for CALICE data. The aim of this work is to reach very high purity for electrons selection in a beam with high contamination of pions. Multivariate data analysis was used to combine shower shape variables and to extract a maximum of available information from the data. A Marlin processor was written that incorporates the ROOT version of TMVA.

1 Introduction

The data used for the multivariate analysis was taken in 2010 in the CERN PS with a mixed beam of electrons, muons, pions and protons with energies from 1-10 GeV. The test beam setup consisted of the high granularity CALICE WHCAL (Tungsten Hadronic Calorimeter) prototype, a data acquisition system, two scintillator trigger stations in front of the calorimeter, two Cherenkov counters upstream for particle selection and three wire chambers used to determine the coordinates of the incident point of the particles on the calorimeter surface. The WHCAL is a sampling calorimeter consisting of 30 layers of sandwich structure. Each layer contains 10 mm thick tungsten absorber, 5 mm thick scintillator tiles and 4 mm iron support. The scintillator tiles have high granularity: 3x3 cm^2 in the center, 6x6 cm^2 and 12x12 cm^2 at the margins. The light from scintillators is transmited through wavelength shifting fibers and read out with Silicon Photomultipliers (SiPM). For more details about the calorimeter, see [1].

The efficiency for electron identification in Cherenkov counters becomes low for low chamber pressure. Therefore it is difficult to separate electrons from pions in the low energy range E<10 GeV. Since electrons and pions show different shower shapes in the calorimeter prototype, multivariate data analysis can be used to optimise the electron/pion separation. The TMVA [2] method called BDT (Boosted Decision Trees) is used for this purpose.

2 Definition of variables

For the classifier training and testing, several Monte Carlo (MC) data sets with 5 GeV input particles were generated. Eleven shower shape variables were used for electron/pion separation (see Fig. 1 to 3):

1. The energy weighted radial distance:

$$d_1 = \frac{\sum E_i \sqrt{(x_i - x_{trk}) + (y_i - y_{trk})}}{\sum E_i}, \tag{1}$$

where E_i is energy of cell, x_i and y_i are the cells center coordinates in the transverse plane, and x_{trk} and y_{trk} are the coordinates of the incident point of the particle on the calorimeter surface.

2. The fraction of the energy contained in the first 5 calorimeter layers E_5/E_{total}, where E_5 is energy sum in the first 5 WHCAL layers and E_{total} is the total energy sum.

3. The third momentum of the radial distance:

$$d_3 = \frac{\sum E_i^3 \sqrt{(x_i - x_{trk}) + (y_i - y_{trk})}}{\sum E_i^3} \tag{2}$$

4. The energy density $\frac{\sum E_i/V_i}{N}$, where V_i is the cell volume and N is the number of cells.

5. The second momentum of the radial distance:

$$d_2 = \frac{\sum E_i^2 \sqrt{(x_i - x_{trk}) + (y_i - y_{trk})}}{\sum E_i^2} \tag{3}$$

6. R_{90}, the radial distance containing 90% of E_{total}.

7. N_{90}/N, the fraction of cells containing 90% of E_{total}.

8. The cells average energy $\frac{\sum E_i}{N}$.

9. L_{max}, the maximum energy loss layer number.

10. L_{start}, the shower start layer number, found with the `PrimaryTrackFinder` [4].

11. $L_{max} - L_{start}$, the number of layers to reach shower maximum.

3 TMVA and BDT

TMVA is a toolkit which hosts a large variety of multivariate classification algorithms. It provides a ROOT- integrated environment for the processing, parallel evaluation and application of multivariate classification and multivariate regression techniques. All multivariate techniques in TMVA belong to the family of "supervised learning" algorithms. They make use of training events, for which the desired output is known, to determine the mapping function that describes a decision boundary (classification). Further information about TMVA can be found in [2].

A decision tree is a structured classifier which takes repeated left/right (yes/no) decisions on a single variable at a time until a stop criterion is fulfilled. This way the phase space is split up in many regions and events are gradually classified as signal or background depending on the final leaf node. Boosting a decision tree extends the concept of a single decision tree. A forest is formed from many trees which are derived from the same training ensemble by reweighting events. Finally, a single classifier is derived which is given by a weighted average of individual decision trees. The advantages of the BDT method are the fact that it is easy to use, not much optimization is needed, it is robust in the presence of correlations and it stabilizes fluctuations in the training sample.

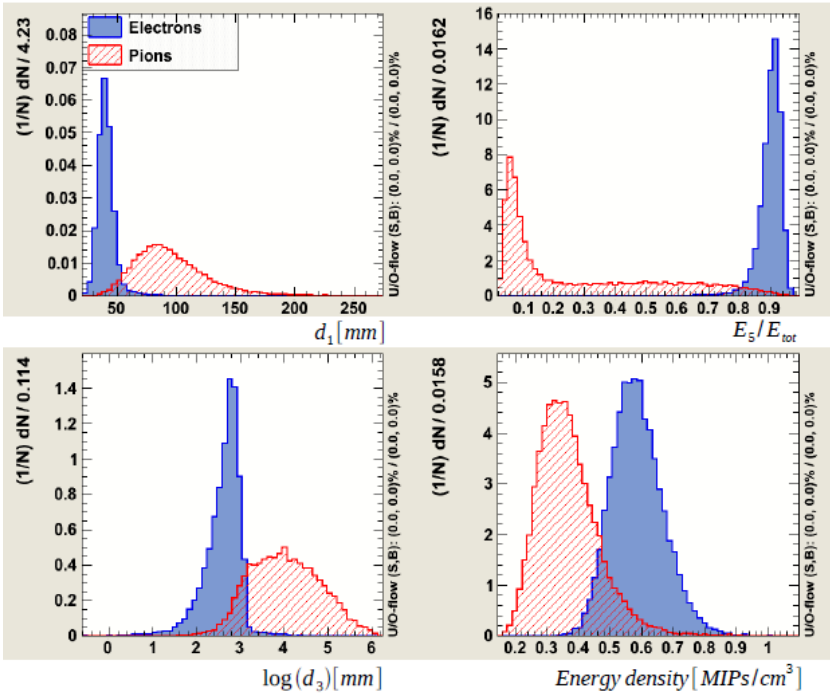

Figure 1: The distributions of the energy weighted radial distance d_1, the fraction of total energy contained in first 5 calorimeter layers E_5/E_{total}, the third momentum of radial distance d_3 and the energy density $\frac{\sum E_i/V_i}{N}$.

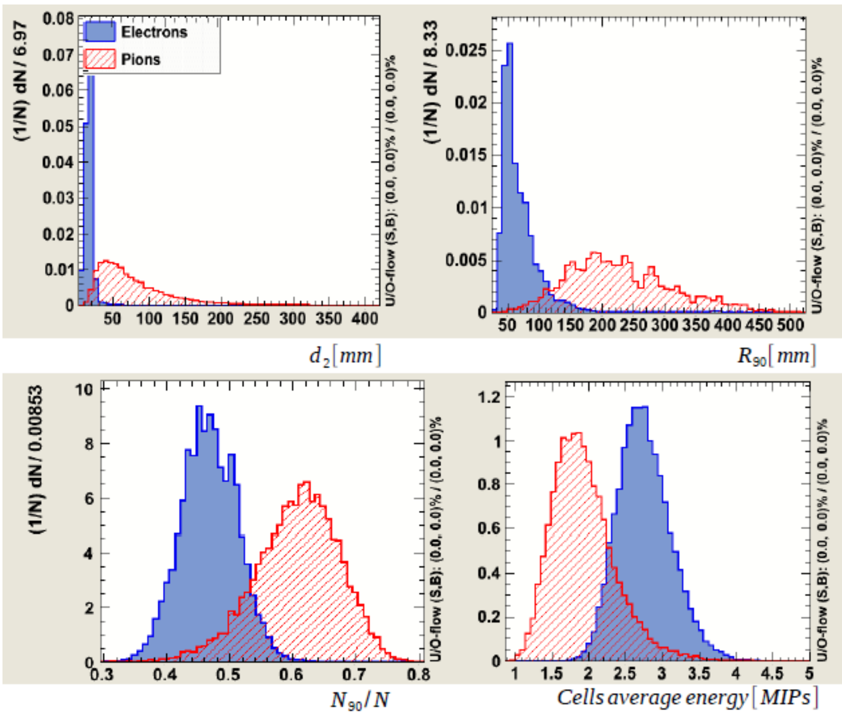

Figure 2: The distributions of the second momentum of radial distance d_2, the radial distance R_{90}, the cells fraction N_{90}/N and the cells average energy $\frac{\sum E_i}{N}$.

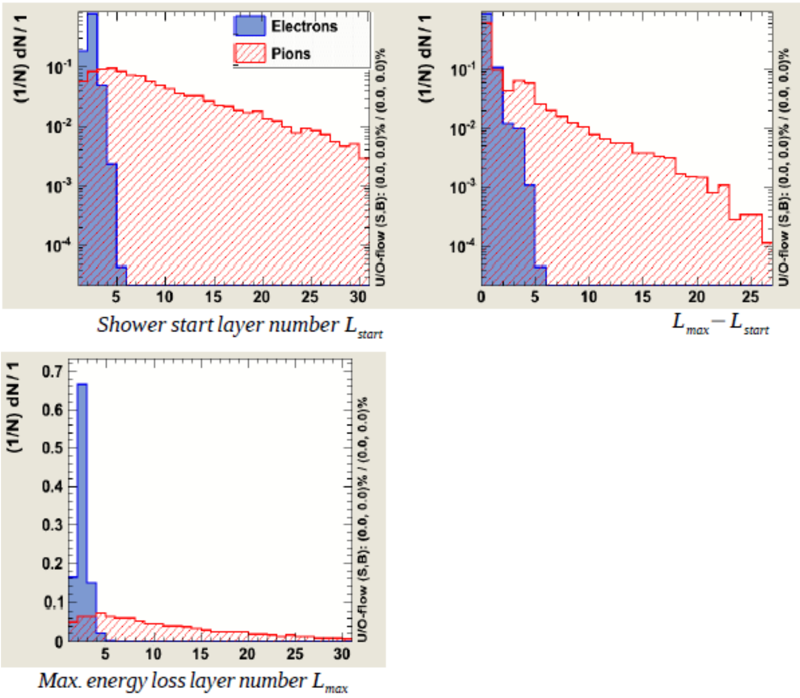

Figure 3: The distributions of the shower start layer number L_{start}, the maximum energy loss layer number L_{max} and the number of layers to reach the shower maximum $L_{max} - L_{start}$.

4 Implementation

A Marlin [3] based package was written for the multivariate analysis of the eleven shower shape variables with the purpose of achieving better electron/pion separation. It consist of:

- Two C++ files:
  - `TMVAClassification.C`: used for filling MC and data ROOT files with shower shape variables information and for classifier training
  - `TMVAClassificationApplication.C`: uses trained classifier for multivariate electron/pion selection.
- Three steering files (in xml format):
  - `tmvaMC.xml`
  - `tmvaData.xml`
  - `tmvaAnalysis.xml`

The first two steering files are used together with `TMVAClassification.C` for creating and filling ROOT files with the eleven variables. The MC ROOT files should be made first avalaible for classifier training. This can be done using the `tmvaMC.XML` steering file. This file has four TMVA parameters: `FillRootTree`, `TMVA_Method`, `TMVA_Training` and `TMVA_Analysis`. The `FillRootTree` parameter tells processor to create ROOT file and fill it with variables information. For this purpose it has to be set to `true`. The remaining three parameters, when creating and filling the ROOT file, have to be set to default values: `TMVA_Method=BDT` (defines classifier method), `TMVA_Training=false` (enables classifier training) and `TMVA_Analysis=false` (enables multivariate electron/pion selection).

The steering file `tmvaData.XML` has the same TMVA parameters as `tmvaMC.XML`. The only difference is that it is used for creating and filling real data ROOT files. While doing this all four parameters should be set to same values like in MC case.

The steering file `tmvaAnalysis.XML` is used with `TMVAClassification.C` for training the BDT classifier with MC ROOT files and with `TMVAClassificationApplication.C` for multivariate electron/pion selection from real data ROOT files. It contains the same parameters as `tmvaMC.XML` and `tmvaData.XML` plus additional parameters for multivariate data analysis. For training and analysis usage, `FillRootTree` should be set to `false`.

When training the classifier:

- `TMVA_Method=BDT`, sets training method.
- `TMVA_Training=true`, enables classifier training.
- `TMVA_Analysis=false`, enables multivariate selection. It should be set to `false` for training.
- `TMVA_Training_SignalFile` is the path to the MC electrons ROOT file for classifier training.
- `TMVA_Training_BackgroundFile` is the path to the MC pions ROOT file for classifier training.
- `TMVA_Analysis_SignalFile` is path to the real data ROOT files for multivariate selection (not used in training).
- `signalWeight` is the signal (electrons) weight for classifier training.
- `backgroundWeight` is the background (pions) weight for classifier training.
- `argumentForPreparingTrainingAndTestTree` sets the TMVA training and testing parameters.
- `argumentForBookMethod` sets the TMVA classifier parameters.

After training the classifier, `TMVAClassificationApplication.C` together with the `tmvaAnalysis.XML` steering file is used for multivariate electron/pion selection. All TMVA parameters remains the same in steering file except:

- `TMVA_Training=false` disables training.
- `TMVA_Analysis=true`, enables multivariate selection.
- `TMVA_Analysis_SignalFile` is set to real data ROOT file path.

When training is finished processor generates `weights.XML` file which is used by `TMVAClassificationApplication.C` for multivariate selection. It also creates the `weights` directory where all classifier training and testing information is stored.

5 Results

For the BDT classifier training and testing, 22587 (22586) MC events were used for the training (testing) of the electron sample, and 18046 (18045) for the training (testing) of the pion sample. The resulting classifier response and the electron/pion cut efficiencies are shown in Fig. 4.

The electron/pion cut efficiencies are defined as follows:

$$eff_{electron} = \frac{N_{electrons\ selected}}{N_{electrons\ total}} \tag{4}$$

$$eff_{pions} = \frac{N_{pions\ selected}}{N_{pions\ total}}. \tag{5}$$

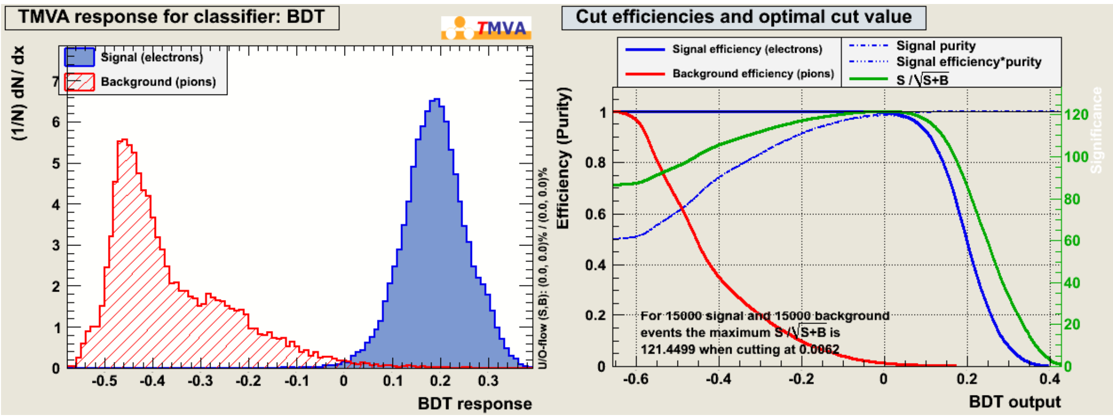

Figure 4: The BDT classifier response (left) and the electron/pion cut efficiencies and optimal cut value (right).

The ranking of the variables used in the decision trees and their importance are given in Table 1. The importance is defined by the number of cuts made for the corresponding variables during the decision making.

Rank	Variable	Importance
1	N_{90}/N	1.288e-01
2	d_1	1.269e-01
3	E_5/E_{tot}	1.165e-01
4	d_2	1.145e-01
5	Energy density	1.100e-01
6	R_{90}	1.018e-01
7	d_3	1.015e-01
8	L_{start}	7.623e-02
9	Cells average energy	6.687e-02
10	L_{max}	4.406e-02
11	$L_{max} - L_{start}$	1.272e-02

Table 1: Variable ranking according to their importance in the decision tree.

In order to assess the performance of the BDT method, the obtained results were compared to the case in which for the electron/pion separation only a simple cut on the two variables was applied: $d_1 \leq 40$ and $E_5/E_{tot} \geq 0.875$ (see Fig. 5).

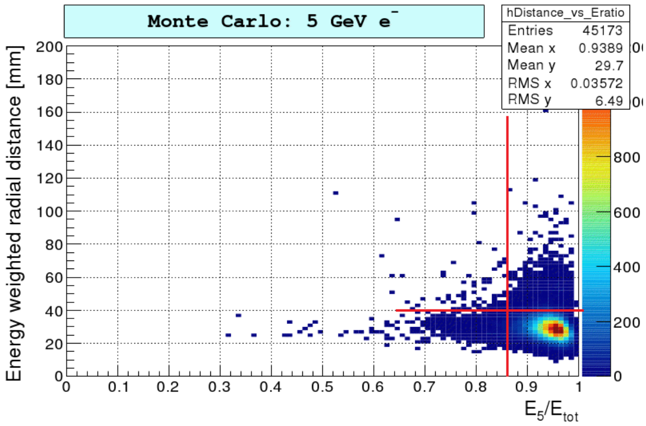

Figure 5: The distribution of the energy weighted radial distance versus the energy fraction in the first five layers. The red lines represent the cut on d_1 $\leq$40 and $E_5/E_{tot} \geq 0.875$.

When applying the simple cut on the two variables, the following efficiencies are obtained: $\varepsilon_{electrons} = 0.48$ and $\varepsilon_{pions} = 0.00047$. Using BDT with eleven input variables and requiring $\varepsilon_{electrons} = 0.48$, we get $\varepsilon_{pions} = 0.00027$ and for $\varepsilon_{pions} = 0.00047$, $\varepsilon_{electrons} = 0.61$. The multivariate selection thus improves both the electron efficiency and the pion suppression compared to the simple cut on two variables done earlier.

TMVA method	Input variables	$\varepsilon_{electrons}$ for $\varepsilon_{pion} = 0.01$	$\varepsilon_{electrons}$ for $\varepsilon_{pion} = 0.1$	Separation $\langle S^2 \rangle$
Optimised cuts	$d_1,\ E_5/E_{tot}$	0.975	0.992	-
BDT	$d_1,\ E_5/E_{tot}$	0.977	1	0.956
BDT	$N_{90}/N,\ d_1,\ E_5/E_{tot}$, d_2, energy density, $R_{90},\ d_3$	0.988	1	0.970
BDT	$N_{90}/N,\ d_1,\ E_5/E_{tot}$, d_2, energy density, R_{90}, d_3, L_{start}, cells average energy, L_{max}, $L_{max} - L_{start}$	0.991	1	0.973

Table 2: Electron/pion efficiencies for the BDT method and for the optimized cut method.

Table 2 compares the performance of the optimised cut method to the one of the BDT method for different sets of input variables. Given are the electron selection efficiencies for pion contaminations of 1 % and 10 %, respectively, as well as the separation $\langle S^2 \rangle$, defined as:

$$\langle S^2 \rangle = \frac{1}{2} \cdot \frac{\int (y_{electron} - y_{pion})^2 \, \mathrm{d}y}{y_{electron} + y_{pion}}, \tag{6}$$

where $y_{electron}$ and y_{pion} are the corresponding probability density functions (PDFs). $\langle S^2 \rangle$ is 0 for full overlap and 1 for no overlap. The electron efficiency for the BDT method with two input variables is slightly higher than for the optimised cut method.

It increases further for a larger number of input variables, with a corresponding increase of the separation.

6 Conclusions

A multivariate data analysis technique was used for electron/pion separation in CALICE WHCAL data. The BDT classifier was trained using an MC ROOT sample containing eleven shower shape variables. The performed multivariate selection with MC data shows that the BDT method allows better separation than the simple cut and than the optimized cut method. An increase in the number of the input variables increases the electron selection for a given pion contamination.

References

[1] The CALICE collaboration, C. Adloff, Y. Karyotakis, J. Repond, A. Brandt, H. Brown, K. De, C. Medina *et al.*, JINST **5**, P05004 (2010). [arXiv:1003.2662 [physics.ins-det]].

[2] `http://tmva.sourceforge.net/`

[3] `http://ilcsoft.desy.de/portal/software_packages/marlin/`

[4] `https://svnsrv.desy.de/websvn/wsvn/General.calice/calice_analysis/trunk/addonProcs/src/PrimaryTrackFinder.cc`