Classification of high-energy antiprotons on electrons background based on calorimeter data in PAMELA experiment

O A Dunaeva$^{1,3,*}$, V V Alekseev$^1$, Yu V Bogomolov$^1$, A D Lukyanov$^1$, V V Malakhov$^2$, A G Mayorov$^{1,2}$ and S A Rodenko$^2$

$^1$ Department of Physics, Yaroslavl State P.G. Demidov University, 150000 Sovietskaya st., 14, Yaroslavl, Russia
$^2$ National Research Nuclear University MEPhI (Moscow Engineering Physics Institute), Kashirskoe highway 31, Moscow, 115409, Russia

E-mail: $^*$olaydy@gmail.com

Abstract. In modern experimental physics a heterogeneous coordinate-sensitive calorimeters are widely used due to their good characteristics and possibilities to obtain a three-dimensional information of particles interactions. Especially it is important at high-energies when electromagnetic or hadron showers are arise. We propose a quit efficient method to identify antiprotons (positrons) with energies more than 10 GeV on electron (proton) background by calorimeter of such kind. We construct the AdaBoost classifier and SVM to separate particles into two classes, different combinations of energy release along reconstructed particle trajectory were used as feature vector. We test a preliminary version of the method on a calorimeter of the PAMELA magnetic spectrometer. For high-energy particles we got a good quality of classification: it lost about $5 \cdot 10^{-2}$ of antiprotons, and less than $4 \cdot 10^{-4}$ of electrons were classified to antiproton class.

1. Introduction

The PAMELA instrument was launched from the Baikonur Cosmodrome on 15 June 2006 on semipolar orbit with inclination 70 degrees onboard the Resurs-DK1 satellite and took data untill January 2016. The instrument was designed to measure the spectra of charged particles in the cosmic rays for an energy interval from hundred MeV to several hundred GeV. Main goal of the experiment was studing antiparticles in the cosmic rays (positrons and antiprotons [1]), since 10 years of observations make PAMELA a good device for such studies. Pamela instrument consists of following detectors: anticoincidence system, a time of flight system, a tracking system, an electromagnetic calorimeter, a neutron detector, shower leak detector S4 [2].

By deflection particle in magnet track system we measure rigidity, charge and sign of charge. But for the separation of particles with the same rigidity such as antiproton and electron we use electromagnetic calorimeter. Calorimeter of PAMELA magnetic spectrometer consists of 44 single-sided silicon sensor planes 38 thick interleaved with 22 plates of tungsten absorber 0.26 cm thick [3]. This corresponds to 0.74 radiation lengths and $\sim 0.6$ nuclear interaction lengths. Each

$^3$ To whom any correspondence should be addressed.
plane consists of 96 strips. The orientation of the strips of two consecutive planes is orthogonal and therefore provides two-dimensional spatial information.

The main task of the calorimeter is to select positrons and antiprotons from protons and electrons and to measure energy of stopped particles electronst and positrons. In figure 1, we can see significant differences in longitudinal and transverse profiles of interactions. For example we calculated $Q_{tot}$ – the total energy release in the calorimeter; $Q_{max}$ – maximum energy release in the calorimeter; $Q_{track}$ – the energy release along the shower axis; $Q_{cyl}(N_{cyl})$ – the energy (number of triggered strips) in the cylinder radius of 4 strips around the shower axis; $Q_{pre}(N_{pre})$ – the energy (number of triggered strips) in the cylinder radius of 8 strips around the axis of the shower in the first three planes.

2. Classification

Identification of high-energy antiprotons on electrons background can be reduced to the classification problem. Suppose we are given $k$ observations of particles. Each observation consists of a pair: a features vector $x_i \in \mathbb{R}^n$, $i = 1, \ldots, k$ and the associated labels $y_i$, given to us by a trusted source. In our task $y_i = +1$ means that $i$-th particle is antiproton, and $y_i = -1$ means that this is electron. We call this collection the training set; it is used to fine-tune parameters of the classifier. Constructed classifier can be used to predict of particle type for examples from a test set. There are some algorithms to construct the classifier, such as neural networks, random forest, boosting, Support Vector Machine(SVM). We consider AdaBoost algorithm and SVM.

The idea of boosting is creating a strong classifier from a set of weak classifiers. A weak classifier $h_t(x)$ is an algorithm which have to be only slightly positively correlated with the correct classification, and strong classifier is a weighted sum of weak classifiers:

$$ H(x) = \text{sign} \left( \sum_{t=1}^{T} \alpha_t h_t(x) \right) . $$

AdaBoost is a special case of gradient boosting with exponential loss function [4]. In our approach we use AdaBoost classifier on Decision Trees.

In 60-70 years a team of mathematicians led by V N Vapnik proposed method generalized portrait, based on the construction of an optimal separating hyperplane. Algorithm require that objects lie from the separating surface as far as possible. In the 90 years the generalized method became known as support vector machines (SVM) [5].
Linear SVM constructs separating hyperplane with the maximum margin (gap between two classes), this task can be found by solving the following optimization problem:

\[
\min \left( \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{l} \xi_i \right)
\]

where \(w\) is the weight vector normal to the hyperplane, and \(\xi_i > 0\) are slack variables for misclassified examples. The parameter \(C\) controls the trade off between errors of the SVM and margin maximization.

Note that in common case classifier algorithms trying to minimize the summary errors and the quality of classification will be similar for both classes. However we want construct the classifier which has percentage of misclassified electrons as small as possible. A lot of classification methods have ability to set weights \(w\) for classes, for example SVM [6]. Thus we can control the quality of electron identification by choosing weights of classes.

3. Data base and results
We simulated antiprotons and electrons spectra similar to cosmic rays \([p^-, e^+]\) [1, 7] in rigidity range from 10 to 1000 GV by software package in Geant4 environment used by PAMELA collaboration. To select good events we applied following criteria: no signal in anticoincidence system; one hit in S1 and S2 of time of flight detector; good track in the tracker (i.e. not close to the magnet, good chi2 etc), absolute value of charge determined by dependence of ionization losses from rigidity; exclude events registered in South Atlantic Anomaly. These criteria allow to reliably determine the direction of the particle arrival, its rigidity, charge, and charge sign.

| particles  | train | test | total |
|------------|-------|------|-------|
| Antiprotons| 24213 | 10378| 34591 |
| Electrons  | 19420 | 8323 | 27743 |

We use cross validation method for testing the quality of trained classifier we obtain on test set. The database was splitted into training and test set in the ratio of 1 to 3 randomly, statistic of database showed in table 3.

| SVM (\(|R| < 1000\)) | antiprotons | electrons |
|------------------------|-------------|-----------|
| antiprotons             | 85.903%     | 14.097%   |
| electrons               | 0.072%      | 99.928%   |

| SVM (\(10 > |R| > 500\)) | antiprotons | electrons |
|-----------------------------|-------------|-----------|
| antiprotons                 | 90.064% (9173) | 9.936% (1012) |
| electrons                   | 0.029% (2)   | 99.971% (6869) |

| AdaBoost (\(|R| < 1000\)) | antiprotons | electrons |
|----------------------------|-------------|-----------|
| antiprotons                 | 92.185%     | 7.815%    |
| electrons                   | 0.072%      | 99.928%   |

| AdaBoost (\(10 > |R| > 500\)) | antiprotons | electrons |
|---------------------------------|-------------|-----------|
| antiprotons          | 96.005% (9128) | 3.995% (386) |
| electrons            | 0.0% (0)     | 100% (6808) |

We test the classifiers quality for different parameters and classes weights. For SVM we have chosen optimally \(C = 0.2\) and \(w_{electron} = 39\), for AdaBoost we have chosen \(w_{electron} = 40\). In
Table 2 presented the best results of SVM and Adaboost classifiers. We can see that results for particle in \([-500 \, GeV, -10 \, GeV]\) is much better, but one test cannot show real situation.

Because we got just few misclassified electrons, the result of classifier depends much on database partition and test set. In table 3 presented statistic of misclassified electrons for 30 random partitions for train and test sets. We can see that results of SVM better then AdaBoost, tight restrictions for rigidity let us to get the best results.

|        | \(R > -1000\) | \(-10 > R > -500\) |
|--------|----------------|---------------------|
|        | mean, \(10^{-4}\) | variance, \(10^{-5}\) | mean, \(10^{-4}\) | variance, \(10^{-5}\) |
| SVM    | 5.551          | 9.299               | 1.208          | 1.185               |
| AdaBoost | 11.891        | 17.963              | 4.380          | 5.418               |
|        | mean, \(10^{-2}\) | variance, \(10^{-4}\) | mean, \(10^{-2}\) | variance, \(10^{-4}\) |
| SVM    | 14.283         | 0.160               | 11.156         | 0.173               |
| AdaBoost | 7.052         | 0.074               | 3.620          | 0.063               |

In assumption that ratio of electrons to antiprotons in primary cosmic rays is about 100 we have about 100 electrons to one antiproton. In case of SVM we got \(1.208 \times 10^{-4}\) of misclassified electrons, it means that to hundred antiprotons we have 1.2 misclassified electrons. But from hundred antiprotons we lost 11.156, thus we have \(1.2 / 88.844 = 0.0135\) impurity electrons in antiprotons. In case of AdaBoost classifier to hundred antiprotons we have 4.38 misclassified electrons. But from hundred antiprotons we lost 3.62, thus we have \(4.38 / 96.38 = 0.0454\) impurity electrons in antiprotons.

4. Conclusion
We showed that machine learning algorithms can be used for identification of antiprotons on electrons background based on calorimeter features and rigidity. SVM with weights of classes showed better rejection of electrons than AdaBoost classifier. Finally we obtained the classifier, which allow to select about 90% - 95% of antiprotons and effectively suppress electrons background on experimental data PAMELA spectrometer. However, we note that presented method needs more statistic for stable classification of antiprotons on experimental data.

Acknowledgments
The research is financed by the Russian Science Foundation grant (Project No. 15-12-10039).

References

[1] Adriani O et al. 2013 Measurement of the flux of primary cosmic ray antiprotons with energies of 60 MeVto 350 GeV in the PAMELA experiment JETP lett. 96(10) 621-2
[2] Picozza P et al. 2007 PAMELA payload for antimatter matter exploration and light-nuclei astrophysics Astroparticle physics 27(4) 296–315
[3] Boezio M et al. 2006 The electron-hadron separation performance of the PAMELA electromagnetic calorimeter Astroparticle physics 26 111–18
[4] Freund Y, Schapire R 1995 A Decision-Theoretic Generalization of on-Line Learning and an Application to Boosting European conference on computational learning theory Springer Berlin Heidelberg 23-37
[5] Burges C 1998 A Tutorial on Support Vector Machines for Pattern Recognition Data Mining and Knowledge Discovery 2(2) 121–67
[6] Batuwita R and Palade V 2013 Class Imbalance Learning Methods for Support Vector Machines Imbalanced Learning: Foundations, Algorithms, and Applications ed He H and Ma Y (NJ: John Wiley and Sons)
[7] Adriani O et al. 2011 Cosmic-ray electron flux measured by the PAMELA experiment between 1 and 625 GeV Physical review lett. 106(20) 201101