



ELSEVIER

Available online at www.sciencedirect.com

SCIENCE @ DIRECT®

Nuclear Instruments and Methods in Physics Research A 525 (2004) 412–416

**NUCLEAR
INSTRUMENTS
& METHODS
IN PHYSICS
RESEARCH**
Section A

www.elsevier.com/locate/nima

Clustering analysis and supervised methods for antiparticle studies in the PAMELA experiment

R. Bellotti^{a,b}, M. Boezio^c, F. Volpe^{a,*},¹

^a*Dipartimento di Fisica, Università di Bari and Istituto Nazionale di Fisica Nucleare, Sezione di Bari, Italy*

^b*Center of Innovative Technologies for Signal Detection and Processing (TIRES), Bari, Italy*

^c*Istituto Nazionale di Fisica Nucleare, Sezione di Trieste, Italy*

Abstract

In this study a new approach to pattern recognition problems in astroparticle physics is presented. The context in which this work has been developed is the satellite borne experiment PAMELA, whose principal aim is antiparticle studies. In particular the classification problem of the PAMELA imaging calorimeter has been taken into account. This detector is designed for particle identification; due to its high granularity, both in the transversal and in the longitudinal direction, the calorimeter is suitable for reconstructing the spatial development of a shower. For each event the calorimeter is able to provide a 3D image that can be used to discriminate between hadrons and leptons. In this work the available information for each kind of image event class has been pre-processed representing each event by means of discriminating variables. A clustering analysis has been applied to a data set and the classification has been performed using supervised algorithms. Results from simulated data from the PAMELA prototype calorimeter will be shown.

© 2004 Elsevier B.V. All rights reserved.

PACS: 98.70.Sa; 07.05.Mh; 95.55.Vj

Keywords: Cosmic ray; Electromagnetic calorimeter; Clustering; Artificial Neural Networks; Support Vector Machines

1. Introduction

Particle identification and separation of signal from background is a very common task for pattern recognition in Astroparticle Physics experiments as well as in High-Energy Physics.

PAMELA is a satellite borne experiment devoted mainly to measure the spectrum of \bar{p} , e^+ and light nuclei in the cosmic radiation in the energy range 10^2 – 10^6 MeV [1]. The PAMELA

telescope will be installed on board the Russian RESURS DK-1 spacecraft and the launch is scheduled for the beginning of 2004.

The PAMELA electromagnetic calorimeter has been designed to perform precise measurements of the total energy released and to reconstruct the spatial development of showers giving a 3D image of each event. The expected rejection power of this detector is 10^4 for p/e^+ and \bar{p}/e^- measurements at a selection efficiency of 95% [2]. The main aim of the PAMELA calorimeter is to distinguish between hadronic and electromagnetic showers by comparing the different topologies of these two kind of events.

*Corresponding author.

E-mail address: francesca.volpe@ba.infn.it (F. Volpe).

¹At present at Royal Institute of Technology, Stockholm.

Traditional methods of data-analysis are based on the study of the distributions of certain descriptive physical variables and on the application of cascade-cuts. In these procedures the choice of the final working point depends on each intermediate cut and presents a certain arbitrariness. In this work we present an innovative approach to pattern recognition problem of the PAMELA calorimeter; a modular classification system has been set up to classify electromagnetic, hadronic showers and non-interacting particles. This system consists first of a clustering phase, in which interacting particles are selected by the dataset; then, in a second step the classification is performed on the surviving data set using supervised algorithms, such as Support Vector Machines or traditional Artificial Neural Networks. The results presented here concern the application of this modular architecture to a simulated dataset obtained from the CERN-GEANT 3.21 official collaboration code GPAMELA [3].

2. Particle selection in PAMELA calorimeter

The calorimeter is composed by 11 modules, each formed by two series of: single-sided silicon plane (Si- X view), tungsten absorber (W), single-sided silicon plane (Si- Y view) for a total number of 44 silicon layers 380 μm thick and 22 absorber layers. In the X - or Y -direction 9 silicon detectors are located to form a square matrix 3×3 and with a total area of $24 \times 24 \text{ cm}^2$.

The calorimeter has a high granularity both in the longitudinal (Z) and in the transversal (X and Y) directions. In the Z -direction the granularity is determined by the thickness of the absorber layers; each tungsten layer is 0.26 cm thick, which corresponds to $0.74 X_0$ (radiation lengths). Since there are 22 tungsten layers, the total depth of the calorimeter is $16.3 X_0$, which is not enough to fully contain the high-energy electromagnetic showers, but is sufficient to allow an accurate topological reconstruction of the shower development. The transverse granularity is due to the segmentation of the silicon detectors into 32 large strips with a pitch of 2.4 mm.

These technical characteristics make the calorimeter a very powerful particle identifier detector:

in fact, it is able to reconstruct the development of the shower providing a 3D image of any event inside the detector.

3. Pre-processing analysis

In the pre-processing phase the whole information produced by the calorimeter is transformed in order to reduce the dimensionality of the problem. Therefore, a number of macroscopic variables were chosen to represent each image-event in the calorimeter. For each particle-event a Region of Interest (RoI) has been selected; the criterion used to select this region consists on searching for the block of 10 consecutive silicon planes having the highest number of hit strips. In this way the RoI can be considered as the longitudinal part of the calorimeter in which most of the physical information from an event is contained. The set of 9 discriminating variables chosen using this criterion are the following: 1. total energy released inside the RoI; 2. total energy released outside the RoI; 3. total number of hit strips inside the RoI; 4. total number of hit strips outside the RoI; 5. total energy released in a cylinder of 1 Moliere radius around the track direction and inside the RoI; 6. total energy released in a cylinder of 1 Moliere radius around the track direction and outside the RoI; 7. total number of hit strips in a cylinder of 1 Moliere radius around the track direction and inside the RoI; 8. total number of hit strips in a cylinder of 1 Moliere radius around the track direction and outside the RoI; 9. total energy released in the plane of maximum interaction, i.e. having the largest energy deposition.

4. Event classification: modular architecture

The first module of the classification system is based on an unsupervised algorithm called Super Paramagnetic Clustering (SPC) [4]. In the first phase of our analysis the clustering algorithm is used for a coarse partitioning of the original data set, according to the different topological characteristics of the pattern produced by the interaction of the particle in the calorimeter. Clustering is

used to determine the non-interacting particles, in order to extract a data set composed only by interacting particles as input to the second phase of the analysis.

In the second phase a supervised method is used for a fine discrimination between electromagnetic and hadronic shower. The supervised algorithms employed to perform the classification are: Support Vector Machines (SVMs) and traditional Artificial Neural Networks (NN).

4.1. Super paramagnetic clustering

Clustering can be formally stated as a problem of partitioning into M groups a number N of patterns, each represented as point \mathbf{x}_i in a d -dimensional metric space.

The *Super Paramagnetic Clustering* algorithm is based on a new approach which exploits the analogy with an inhomogeneous Potts ferromagnet to map stable cluster partitions in thermodynamic phases of the ferromagnetic system. A Potts spin variable s_i having q states is assigned at each point \mathbf{x}_i (that represents one of the patterns) and a short range ferromagnetic interaction J_{ij} is introduced between pairs of neighboring spins, whose strength decrease as the inter spin distance $\|\mathbf{x}_i - \mathbf{x}_j\|$ increases. The system is governed by the following Hamiltonian (energy function)

$$H = \sum_{\langle i,j \rangle} J_{ij}(1 - \delta_{s_i s_j}) \quad s_i = 1, \dots, q \quad (1)$$

and exhibits three phases. At very low temperatures it is completely ordered: i.e. all the spins are aligned (*ferromagnetic phase*). At very high temperatures the system does not exhibit any ordering and all the spin-pairs have no correlation; this phase is called *paramagnetic*. At intermediate temperatures the global order of the very low temperatures disappears, while regions characterized by a local ordering naturally emerge. Strongly coupled spins within the same high density region become completely aligned, while different regions remain uncorrelated. This intermediate phase is called *super paramagnetic* and can be imagined as an ensemble of magnetic grains (i.e. *clusters*).

By means of a Monte Carlo, the SPC algorithm simulates an ensemble of Potts-spin configurations

and calculates the thermal average of some variables associated with each spin configuration, such as the average magnetization. The spin–spin correlation function (the average of $\delta_{s_i s_j}$) is the probability that two spins are aligned and it is used to assign the spins to each cluster. The temperature T is the resolution parameter when searching for the superparamagnetic phase, in which, since clusters of spins behave like independent ferromagnetic domains, clusters are identified as aligned spin domains.

4.2. Supervised algorithms: Support Vector Machines and Neural Networks

SVMs [5] are sophisticated and powerful non-parametric classifiers, with many different configurations. Indeed, many function can be used for the mapping the original training data into a high dimensional feature space in which the decision boundary is determined. SVMs minimize *structural risk* [5], i.e. instead of constructing a decision function f by minimizing the training error on a representative data set (as NN do), it is chosen in a way to minimize an upper bound on the test error. The minimization of the *structural risk* allow an increment of the confidence level with which the classifier classifies unseen data set and consequently its generalization power.

As comparison with the SVMs, a three-layered feed-forward Neural Network [6] has been considered. The network used for this work consists of: nine input neurons (each one associated with a physical variable enumerated in the previous section); a hidden layer composed by five neurons and one output neuron. As transfer function a sigmoid has been used and in the training phase the weights w_{ij} on the network have been iteratively updated by the gradient descent rule:

$$\Delta w_{ij}^{\text{new}} = -\eta \frac{\partial E}{\partial w_{ij}} + \alpha \Delta w_{ij}^{\text{old}} \quad (2)$$

where E is the error on the network outputs y_i with respect to the targets $O_i = 0, 1$

$$E = \frac{1}{2} \sum_{i=1}^N (O_i - y_i)^2. \quad (3)$$

5. Results and conclusions

This modular architecture has been applied to classify 2×10^4 simulated e^- and 2×10^4 simulated π^- with momentum 40 GeV/c. The output of the first step clustering analysis is shown in Fig. 1; it reveals a range of temperatures ($0.03 < T < 0.055$) in which the clustering solution is stable. In this range the classes individuated are 4. The dataset is considered clusterized at the value $T = T^*$, at which the transition to the superparamagnetic phase occurs, i.e. corresponding to the change of the magnetization's variance [4]. In our case $T^* = 0.485$ and at this value the composition of the classes individuated by SPC methods is reported in Table 1. The four classes can be tagged on the basis of the statistical distribution of the nine macroscopic physical variables. In our case it results that classes number 2 and 4 corresponds to high purity non-interacting π^- , i.e. π^- which do not produce any shower, whereas classes number 1 and 3 are interacting particles.

The interacting particles individuated and tagged by the SPC has been sent as input to the supervised algorithms of the second step of this modular analysis and the classification between hadronic and electromagnetic showers has been performed. In Fig. 2 the performances obtained

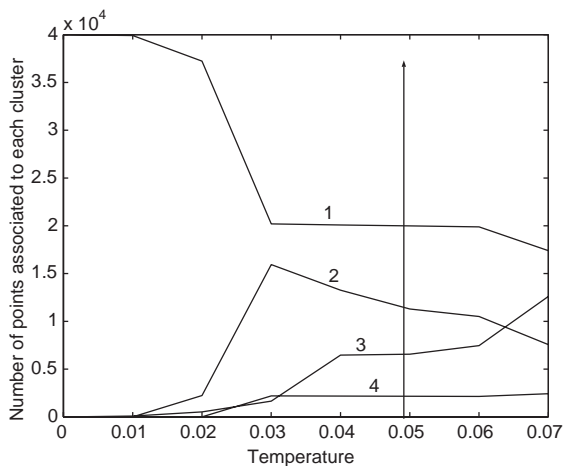


Fig. 1. Output of SPC; the arrow corresponds to the value $T^* = 0.485$ at which the transition to superparamagnetic phase occurs.

Table 1
Number of events in each class individuated by SPC at $T = 0.485$

| Class | Number of events per class |
|-------|----------------------------|
| 1 | 19993 |
| 2 | 11290 |
| 3 | 6559 |
| 4 | 2158 |

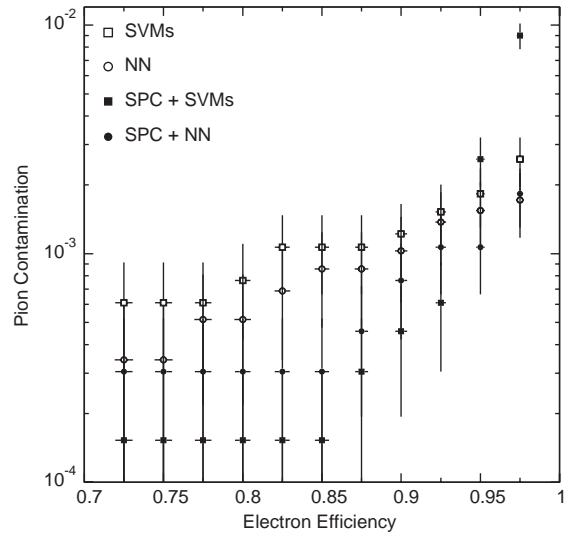


Fig. 2. Pion contamination vs electron efficiency using the modular SPC + SVMs or SPC + NN and stand alone SVMs or NN.

with both the modular systems, i.e. SPC + SVMs and SPC + NN, are shown. In comparison, performances obtained applying stand alone supervised algorithms (SVMs and NN) are reported.

In this paper we have presented a modular system of PAMELA calorimeter data-analysis, based on clustering techniques combined with two supervised algorithms. Classification performances obtained by this modular system are around 10^{-4} in pion contamination at an electron efficiency above 80%, although a larger data set is needed for a more precise study. These results outperform those obtained by stand alone supervised algorithms. Moreover, this modular techniques allows an accurate estimate of the rejection/efficiency curve. This is particularly important because during the experiment life time making it necessary

to choose a different calorimeter working point for the off-line analysis.

References

- [1] V. Bonvicini, et al., Nucl. Instr. and Meth. A 461 (2001) 262.
- [2] M. Boezio, et al., Nucl. Instr. and Meth. A 487 (2002) 407.
- [3] <http://www.ba.infn.it/~ambriola/gpamela>.
- [4] E. Domany, et al., Phys. Rev. Lett. 76 (1996) 3251.
- [5] V. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.
- [6] C.M. Bishop, Neural Networks for Pattern Recognition, Oxford Press University, Oxford, 1995.