Available online at www.sciencedirect.com
SCIENCE DIRECT®
Nuclear Instruments and Methods in Physics Research A 559 (2006) 129–133
NUCLEAR INSTRUMENTS & METHODS IN PHYSICS RESEARCH Section A
www.elsevier.com/locate/nima

A segmented principal component analysis applied to calorimetry information at ATLAS

H.P. Lima Jr.[a,b,*], J.M. Seixas[b]

[a]Brazilian Center for Research in Physics, Rua Dr. Xavier Sigaud 150, Rio de Janeiro 22290-180, Brazil
[b]Signal Processing Laboratory, COPPE/EP-UFRJ, CP 68504, Rio de Janeiro 21945-970, Brazil

Available online 15 December 2005

Abstract

A segmented principal component analysis is applied for dimensionality reduction of the calorimeter information at the second level trigger of ATLAS. The segmented analysis is proposed in order to fully explore the high segmentation of the calorimeter system and the different levels of granularity present at each segment of the hadronic and electromagnetic sections. Considering electron and jet simulated events, a high data compaction level (above 96%) is achieved, even when preserving 95% of the original data variance. Using data projection onto the principal components of each calorimeter segment, and a neural classifier, 97.3% of electrons are correctly identified for a misclassification of jets below 9%.

PACS: 07.05.Kf; 07.05.Mh; 29.40.Vj

Keywords: Principal component analysis; Calorimeters; Particle identification

1. Introduction

The Large Hadron Collider (LHC) will be colliding two bunches of protons at every 25 ns, producing a huge amount of data to be processed. Data comprise both the physics of interest, such as the signatures of the Higgs boson, and a deep background noise. In this scenario, complex trigger systems need to be designed in order to select only the interesting events.

The ATLAS trigger system consists of three distinct levels of event selection [1]. Each trigger level should perform specific algorithms to select only the events with high probability in carrying interesting physics information. From an initial bunch crossing rate of 40 MHz, the ATLAS trigger system will select events up to 100 Hz to permanent storage. The first level trigger looks at detector data with reduced granularity in order to take a fast decision, delivering events to the second level at a maximum rate of 100 kHz. At the second level, complex algorithms operate with the full granularity of the detector, guided by *regions of interest* (RoIs), which contain interesting features of the events [2]. This second level reduces the event rate to less than 1 kHz. The last step of data selection, the Event Filter, performs even more complex algorithms to reduce further the event rate to a maximum of 100 Hz, which corresponds to the data to be permanently stored for offline analysis. The three levels of selection make use of the information provided by the calorimeter system of ATLAS, due to the fast response of the detectors and the ability in identifying particles from the energy deposition patterns [3].

The online triggering system requires both fast signal processing and high efficiency in event selection. For detailed description of particle interactions, the calorimeter system in ATLAS is segmented into four layers in the electromagnetic section and three layers in the hadronic one (see Table 1). In addition, the number of cells and the corresponding granularity vary according to data RoI. Therefore, not only the event selection task at the second

*Corresponding author. Brazilian Center for Research in Physics, Rua Dr, Xavier Sigaud 150, Rio de Janeiro, 22290-180, Brazil.
E-mail addresses: herman@lps.ufrj.br, hlima@cbpf.br (H.P. Lima Jr.), seixas@lps.ufrj.br (J.M. Seixas).

0168-9002/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.nima.2005.11.131

Table 1
Data compaction level achieved for each sub-detector layer

Sub-detector	Original dimension	85%		90%		95%	
		serpentine	*ring*	*serpentine*	*ring*	*serpentine*	*ring*
Pre-sampler (barrel/endcap)	165	22	3	27	4	36	11
EM Calo (barrel/endcap)—front layer	1520	210	8	250	17	312	44
EM Calo (barrel/endcap)—middle layer	800	52	3	69	5	103	10
EM Calo (barrel/endcap)—back layer	400	25	2	33	4	54	10
Hadronic Calo (barrel)—layer 0	100	19	10	21	15	25	23
Hadronic Calo (barrel)—layer 1	90	10	2	12	3	15	8
Hadronic Calo (barrel)—layer 2	40	3	1	3	1	3	1
Total	3115	341	29	415	49	548	107

level trigger is quite difficult, but also processing speed suffers from this extremely high input data dimensionality. As a consequence, an efficient data compaction scheme is mandatory.

This paper proposes the use of Segmented Principal Component Analysis (SPCA) for data compaction and neural classification using projected data. The SPCA provides data representation at the layer level, instead of global random process representation, which is the aim of Principal Component Analysis (PCA) [4]. Thus, the highly segmented calorimeter information can be fully explored after projecting RoI data onto the principal components of each calorimeter layer. Due to the high data compaction rate of SPCA, a compact neural classifier can be designed for achieving efficient and fast event selection.

2. Proposed system

The proposed signal processing will operate at Level 2, on calorimeter data, in order to

(1) reduce the high processing load due to the high granularity of the information;
(2) speed up the particle identification process;
(3) achieve high particle identification efficiency by means of relevant feature extraction.

The signal processing chain comprises three building blocks: data assembling, data projection and neural classification, as illustrated in Fig. 1. The first step consists of assembling raw data coming from both hadronic and electromagnetic calorimeters. For the specific database used in this work, raw data comprise simulated events of dimension up to 3115 (calorimeter cells). As PCA is based on searching for directions that account for maximum data variance, data assembling is required to format incoming events in an adequate form to the feature extraction process. The formatted output events are row vectors of a fixed dimension, for which each element corresponds to the energy deposited by the incident particle in a calorimeter cell. The second functional block, *data projection*, receives

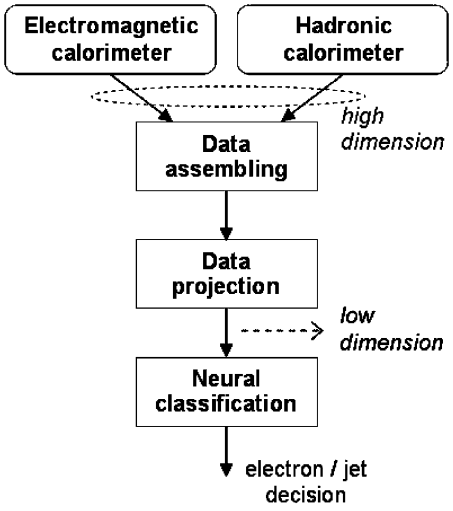

Fig. 1. Signal processing chain.

formatted events and projects them into the principal component space. The projection makes use of eigenvectors previously computed from formatted data of a developing data set. Note that the offline extraction is made separately for each sub-detector (calorimeter layer), as will be described in Section 2.2. Only the eigenvectors with the largest eigenvalues are retained to form a matrix of linear transformation vectors. These vectors are used to generate data projections in the reduced dimension space. Finally, these compacted events are concatenated to feed a neural network that is designed to perform electron/jet separation.

2.1. Data assembling

Simulated LVL2 data produced in the Athena environment [5] were used to test the proposed system. Data correspond to jets and two signatures of the Higgs boson, produced at low luminosity ($L = 10^{33}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$), in the following decays: $H \to 2e^{-}2\mu$ and $H \to 4e^{-}$. The Higgs events have 130 GeV in the center of mass and the electrons are 20 GeV or 30 GeV events.

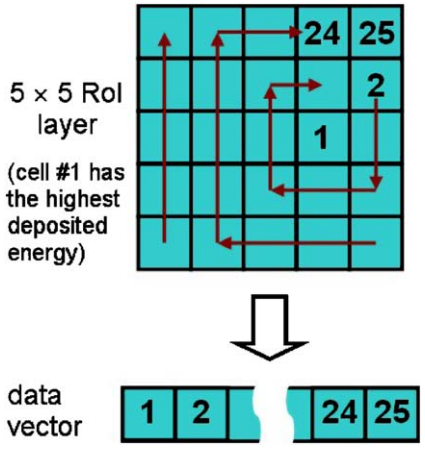

Fig. 2. Ring assembling. See text.

The simulated raw data correspond to interactions with seven sub-detector layers from both electromagnetic and hadronic calorimeters. These sub-detectors are: pre-sampler (barrel and endcap), EM calorimeter (barrel and endcap) front, middle and back layers and Hadronic calorimeter (barrel) layers 0, 1 and 2.

Two approaches for data assembling have been tested, named *serpentine* and *ring*. In *serpentine* assembling, each data vector is a serpentine concatenation of cells in the way they appear in the RoI layer. This means that no topological reorganization is performed on raw data. Due to variability in dimension and missing cells occurring, in practice, each incomplete RoI is filled in with zeros appended to the end in order to keep a pre-defined fixed length. Therefore, this assembling scheme has the drawback of increasing artificially the original data variance, as it is clear from the results shown in Table 1. In the *ring* approach, for each calorimeter layer, the cell with the highest deposited energy is identified, and the data vector is formed by sequentially grouping rings of cells around this marked cell. Fig. 2 illustrates an example with an hypothetic 25-cell RoI. This type of assembling puts in evidence the energy deposition profile of the incident particle, which is an important feature that makes further classification easier to achieve [6]. Moreover, it optimizes data variance description by minimizing the data variance increase when zeros are eventually added.

2.2. Data compaction

In order to perform data compaction, the use of PCA in a segmented way is proposed. This statistical technique, which is based on the *Karhunen–Löeve transformation*, is widely used in multivariate analysis of random processes, being very useful for dimensionality reduction and for finding patterns in data of high dimension [7].

The basic idea in PCA lies in the eigenstructure of the data covariance matrix. Actually, it has been shown that this matrix (considering a random process $\mathbf{x}$ with zero mean ($E[\mathbf{x}] = 0$)) contains the principal directions along which the variance of data has the extremal values. The associated eigenvalues define these extremal values. In short, if the principal directions are represented by $\mathbf{u}_j$, we can define a linear orthogonal transformation of data $\mathbf{x}$ as

$$a_j = \mathbf{u}_j^{\mathrm{T}}\mathbf{x} = \mathbf{x}^T\mathbf{u}_j, \quad j = 0, 1, \ldots, p-1 \tag{1}$$

where a_j are the projections of $\mathbf{x}$ onto the principal directions. The a_j are called the *principal components* and have the same physical dimension of the data vector $\mathbf{x}$. In order to reduce the original data dimension, one may use only the major p projections in Eq. (1), discarding the projections of smaller variance.

PCA was applied to each sub-detector layer separately to better explore the segmented structure of the calorimeter system [8]. The entire simulated data set (22 581 electrons and 7509 jets) was randomly splitted into developing (*training*) and *test* sets. Principal components were extracted only from the training data set, which comprises 11 283 electrons and 3735 jets. Table 1 illustrates the level of compaction achieved for each sub-detector layer, for three different levels of random process energy preservation, 85%, 90% and 95%. As expected, SPCA achieved higher compaction levels on *ring* assembling data than on *serpentine* data. This is due to the better exploration of the energy deposition profile in the first approach. Fig. 3 illustrates the relation between the principal components and data variance for the electromagnetic and hadronic calorimeters.

2.3. Particle identification

Electron/jet classification is performed by a feedforward fully-connected neural network [9]. After data projection onto the reduced sub-space, vectors are concatenated to form a unique data vector, which is then normalized and fed into the neural classifier. Normalization is needed in order to accommodate data values within the dynamic range of the activating functions of the neural network. The adopted normalization was to divide each data vector by its maximum value.

The neural classifier has a three-layer structure with one input layer of source nodes, one hidden layer (10 neurons) and an output layer (1 neuron). Each neuron uses the hyperbolic tangent as the activation function. Network training was performed using the Resilient Backpropagation (RPROP) learning algorithm [10]. This algorithm eliminates the harmful effects of the magnitudes of the partial derivatives. Only the sign of the derivative is used to determine the direction of the weight update. Training runs were carried out with the training data set, already employed for principal component extraction, using test data for validation. Each training step comprised a random selection of an electron/jet pair, in order to reduce overtraining on electrons due to the difference in statistics.

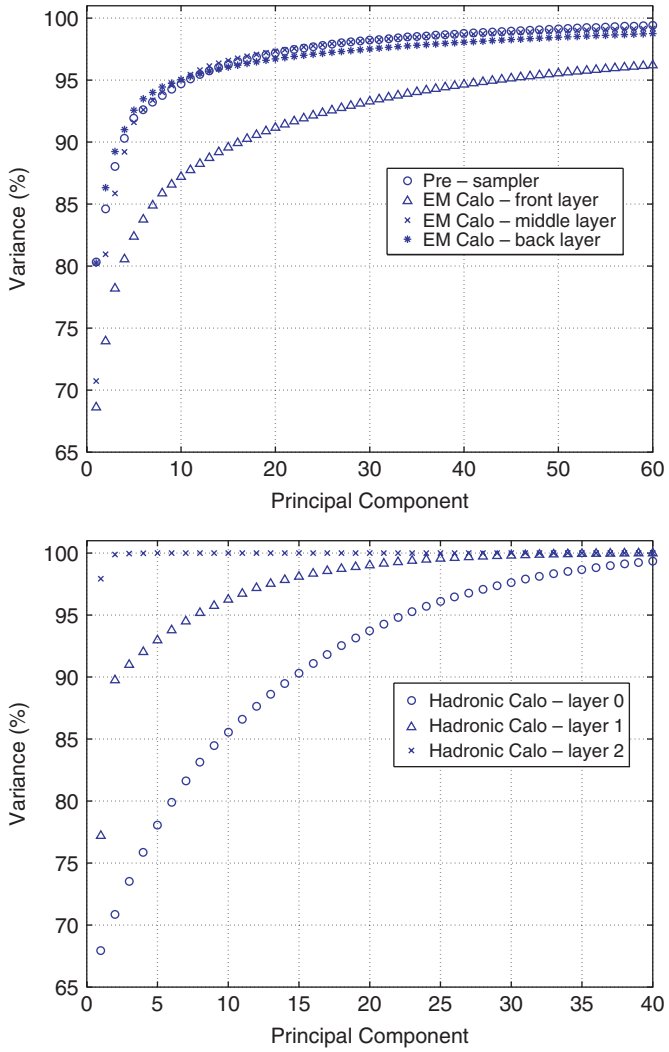

Fig. 3. Load curves for the electromagnetic (top) and hadronic (bottom) calorimeters.

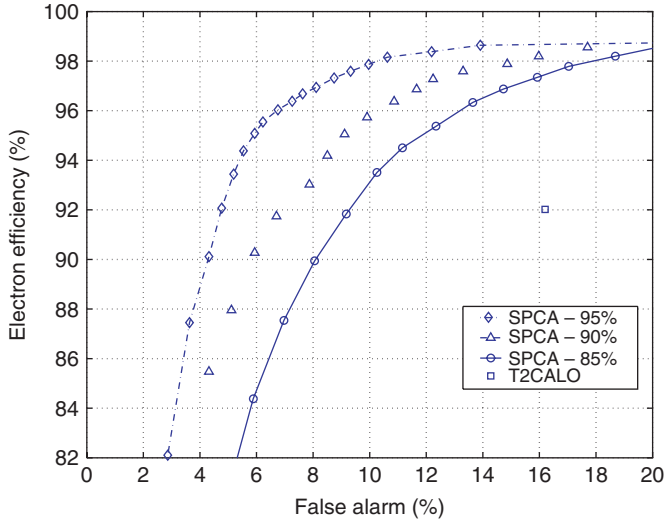

Fig. 4. Receiver Operating Characteristic (ROC) for SPCA and T2Calo. See text.

Fig. 4 compares the Receiver Operating Characteristic (ROC) curves [11] for three levels of data compaction (energy preservation of 85%, 90% and 95%) using SPCA data with *ring* formatting. Results from baseline discrimination algorithm used in ATLAS (T2Calo) [12] are also shown. For an electron detection efficiency of 92%, the proposed discriminator (95% energy preservation) achieves a false alarm probability of 4.8%, whereas T2Calo presents 16.2%.

3. Conclusions

A new particle discrimination scheme, based on segmented principal component analysis and neural networks, was proposed for electron/jet separation at the second level trigger of ATLAS. The extraction of the principal components is performed on each sub-detector layer, exploring the highly segmented structure of the calorimeters. Results from this extraction scheme demonstrate that preserving 95% of the original random process energy, raw events of dimension 3115 may be compacted to 107 principal components. By feeding a neural classifier with these compacted data, 97.3% of electron efficiency was achieved for a false alarm probability of 8.7%. These numbers outperform results from the baseline discriminator in use.

The relevance of each principal component is under investigation. This study may increase the classification efficiency by revealing components with low energy representation but highly discriminant. The use of Non-linear PCA [13] is also being considered in order to achieve even higher compaction levels. Different normalization schemes are under study envisaging to improve the signal-to-noise ratio of data representation.

Acknowledgements

We are thankful for the support from CNPq, CAPES and FAPERJ, in Brazil. We are also thankful to the Trigger/DAQ collaboration for providing the simulation data and for fruitful discussions concerning this work.

References

[1] Level-1 trigger, Technical Design Report, ATLAS TDR-12, 1998.
[2] High-level trigger, data acquisition and controls, Technical Design Report, ATLAS TDR-016, 2003.
[3] R. Wigmans, Calorimetry: Energy Measurement in Particle Physics, Oxford University Press, Oxford, 2000.
[4] I.T. Jolliffe, Principal Component Analysis, Springer, New York, 1986.
[5] Athena, The ATLAS Common Framework, User Guide and Tutorial, Ver. 2, Ed. 2, Draft, CERN, 2001.
[6] A. dos Anjos, J.M. de Seixas, Nucl. Instr. and Meth. A 502 (2003) 713.
[7] K. Mardia, J. Kent, J. Bibby, Multivariate Analysis, Academic Press, New York, 1980.
[8] M.R. Vassali, J.M. de Seixas, Principal component analysis for neural electron/jet discrimination in highly segmented calorimeters,

Proceedings of the VII International Workshop on Advanced Computing and Analysis Techniques in Physics Research, 2000.

[9] S. Haykin, Neural Networks: A Comprehensive Foundation, second ed., Prentice-Hall, Englewood Cliffs, NJ, 1999.

[10] M. Riedmiller, H. Braun, A direct adaptive method for faster backpropagation learning: the RPROP algorithm, Proceedings of the IEEE International Conference on Neural Networks, 1993, pp. 586–591.

[11] H.L. Van Trees, Detection, Estimation and Modulation Theory, Part I, reprint ed., Wiley-Interscience, New York, 2001.

[12] S. Gonzalez, et al., Selection of high-pT electromagnetic clusters by the second level trigger of ATLAS, ALTAS Internal Note, ATL-DAQ-2000-002, 2000.

[13] M.A. Kramer, AIChE J. 37 (2) (1991) 233–243.