ANÁLISE DE COMPONENTES INDEPENDENTES PARA FILTRAGEM ONLINE NUM AMBIENTE DE ALTA TAXA DE EVENTOS E INFORMAÇÃO SEGMENTADA

Eduardo F. Simas F.*, José Manoel de Seixas*, Luiz Pereira Calôba*

*Laboratório de Processamento de Sinais, COPPE/Poli-UFRJ
Rio de Janeiro-RJ, 21945-970.

Emails: esimas@lps.ufrj.br, seixas@lps.ufrj.br, caloba@lps.ufrj.br

Abstract— An online particle discrimination system is proposed for the high level trigger of ATLAS particle detector, which will be placed at one of the collision points of LHC, the next generation particle collider experiment. For feature extraction, Independent Component Analysis (ICA) is applied over a highly segmented calorimeter (energy measurement system). As a pre-processing step the high dimension signals were compacted through Principal Component Analysis (PCA). The hypothesis testing is performed by a neural classifier operating over the independent componentes. After an analysis of the classifier inputs relevance, discrimination efficiency of 97% was achieved for a false alarm probability of 2.5%.

Keywords— Independent component analysis, neural classifiers, detectors, caloremeters.

Resumo— Um sistema online de discriminação de partículas é proposto para o alto nível de filtragem do detector ATLAS, que estará posicionado em um dos pontos de colisão do acelerador de última geração LHC. Para extração de características, Análise de Componentes Independentes (ICA) é aplicada sobre sinais de um calorímetro (sistema de medição de energia) altamente segmentado. Análise de Componentes Principais (PCA) foi utilizada para compactar os sinais de alta dimensão. O teste de hipótese é realizado por um classificador neural operando sobre as componentes independentes. Após um estudo da relevância das entradas do classificador, foi obtida uma eficiência de discriminação de 97% para um falso alarme de 2,5%.

Keywords— Análise de componentes independentes, classificadores neurais, detectores, calorímetros.

1 Introdução

O detector ATLAS (*A Toroidal LHC Apparatus*) objetiva a pré-seleção e armazenamento dos eventos de interesse gerados nas colisões do acelerador de partículas LHC (*Large Hadron Collider*)(Price, 2002). O LHC entrará em funcionamento no segundo semestre de 2007 no CERN (Centro Europeu para Pesquisa Nuclear) e irá colidir feixes de prótons a cada 25ns. Embora produza uma alta taxa de eventos, os canais de interesse no LHC são raros. Um sistema eficiente de filtragem online deve ser projetado para garantir que a maior parte do ruído de fundo seja eliminado, sem desprezar informações valiosas.

Considerando que, em média, cada evento do LHC gera 1,5 Mbytes de informação, é produzido no detector aproximadamente 60 $Tbyte/s$. Mesmo que essa quantidade de informação pudesse ser armazenada, a seleção offline de eventos seria proibitivamente lenta, então, o processo de filtragem de eventos deve ser realizado online, sob severas restrições de tempo de processamento.

O propósito dos experimentos em física de partículas de altas energias é verificar a validade dos modelos teóricos e buscar a solução para problemas ainda sem resposta. No LHC, um dos principais desafios é identificar o bóson de Higgs, que, se existir de acordo com o Modelo Padrão, será responsável por fornecer massa às partículas (Martin, 2006).

Localizado em um dos pontos de colisão do LHC, o detector ATLAS foi projetado para cobrir um ângulo sólido próximo a 4π em torno do ponto de colisão, tendo formato cilíndrico com o túnel do LHC como eixo central (ATLAS-Collaboration, 1999). A Figura 1 mostra um corte transversal do ATLAS com os seus principais

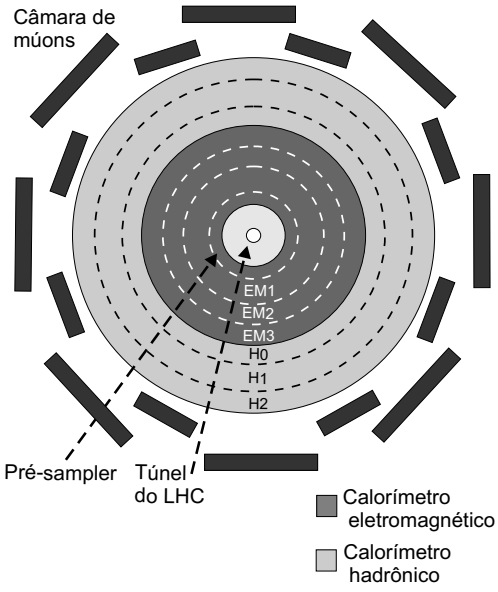

Figura 1: Corte do ATLAS mostrando a segmentação dos calorímetros.

sub-sistemas. Em destaque estão os calorímetros eletromagnético e hadrônico, objetivos deste trabalho, que são segmentados em, respectivamente, 4 e 3 camadas. Os calorimetros medem o perfil de energia depositado pelas partículas incidentes, in-

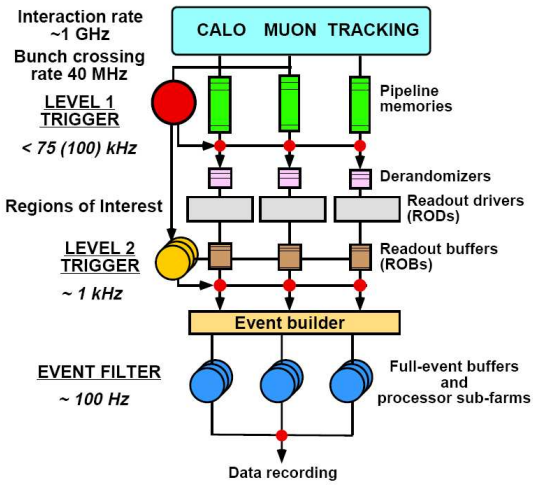

Figura 2: Diagrama em blocos do sistema de filtragem e aquisição de dados do ATLAS, extraído de [1].

formação importante que guia o processo de identificação das partículas. Para tal, a granularidade varia entre as diferentes camadas do calorímetro.

Conforme diagrama da Figura 2, o sistema de filtragem do ATLAS é baseado em 3 níveis seqüenciais de seleção de eventos (ATLAS-Collaboration, 1999). O primeiro nível (LVL1) tem sérias restrições quanto ao tempo de processamento, recebendo a plena taxa de eventos do LHC como entrada. Esse nível é implementado em hardware dedicado, e opera com granularidade menos fina no calorímetro. O LVL1 entrega ao segundo nível a localização das áreas onde possivelmente aconteceram eventos de interesse, regiões estas conhecidas como RoI (*Regions of Interest*). O segundo nível e o filtro de eventos, compõem a filtragem alto nível (HLT - *High Level Trigger*).

A seleção de eventos no HLT é feita através de software especializado, rodando em um conjunto de 2000 PCs (duplos) dedicados operando em paralelo. Nesse estágio a granularidade total dos calorímetros está disponível. A estratégia de processamento seqüencial, adotada no sistema de filtragem do ATLAS, permite que os eventos sejam rejeitados na primeira etapa de análise online possível, minimizando a necessidade de acesso a informações, e facilitando o ajuste e eventual modificação das estratégias de extração de características.

Dentre os eventos gerados nas colisões do LHC, apenas uma pequena parte será útil para a caracterização dos processos da nova física. Por exemplo, assinaturas de interesse podem ser encontradas a partir do decaimento de elétrons, fótons e múons. Os elétrons estão imersos em um intenso ruído de fundo composto de jatos hadrônicos. Os jatos, presentes no alto nível de filtragem, apresentam perfil de deposição de energia nos calorímetros semelhante ao de elétrons (altamente concentrados no calorímetro eletromagnético e quase nenhuma energia depositada na seção hadrônica). O canal elétron/jato (e^-/j) será um dos mais freqüentes no ATLAS (ATLAS-Collaboration, 2003), sendo portanto, muito importante para o desempenho geral do detector.

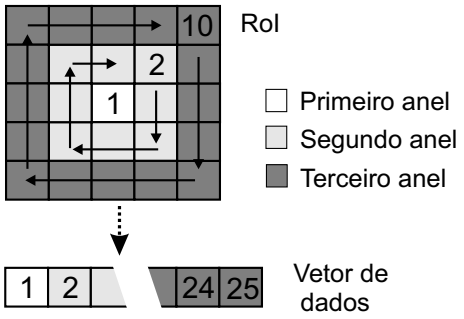

Figura 3: Formatação dos sinais do calorímetro em anéis.

Conforme proposto em (Anjos et al., 2006), neste trabalho os sinais das RoI são formatados em anéis concêntricos de deposição de energia organizados por camada. A célula que absorveu maior energia em uma dada camada é considerada como o primeiro anel, e os anéis subseqüentes são formados pelas células que envolvem o anel anterior (ver Figura 6). Em seguida, as leituras de energia das células de cada anel são somadas e normalizadas por camada do calorímetro. Esse procedimento torna os sinais independentes do ponto de colisão. Assumindo que a dimensão das RoI é fixa e engloba a informação relevante para a discriminação das partículas, são considerados um total de 100 anéis. Sinais típicos de elétrons e jatos na formatação utilizada são mostrados nas Figuras 4 e 5.

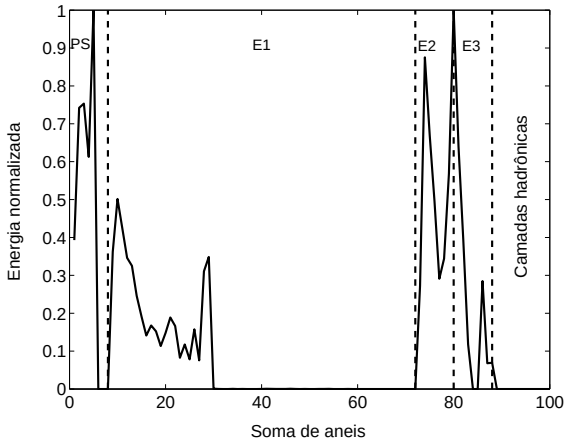

Figura 4: Típico sinal formatado em anéis de um elétron.

O processo de discriminação de partículas pode ser dividido em dois impotantes procedimentos, a extração de características (onde as infor-

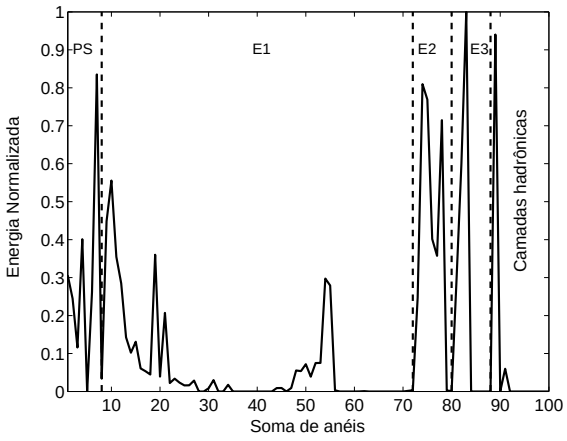

Figura 5: Típico sinal formatado em anéis de um jato.

mações do detector são processadas numa tentativa de facilitar o processo de classificação) e o teste de hipóteses (rotina que realiza a separação das partículas). Neste trabalho é proposta uma nova estratégia para a extração de características, na filtragem de alto-nível do ATLAS no canal elétron/jato, utilizando Análise de Componentes Independentes (ICA - *Independent Component Analysis*). Como pré-processamento, os sinais formatados em anéis (de alta dimensão) são compactados através de Análise de Componentes Principais (PCA - *Principal Component Analysis*). Após a extração de características, classificadores neurais são utilizados para o teste de hipóteses. O objetivo é obter uma melhor performance de discriminação de elétrons e jatos quando comparado ao T2Calo, algoritmo atualmente adotado pela Colaboração do ATLAS e pronto para ser implementado na filtragem de alto nível (ATLAS-Collaboration, 2003). O tempo total de processamento do algoritmo proposto deve ainda atender as limitações impostas para o HLT.

2 Extração de Características

Na Figura 6 é mostrado um diagrama da técnica proposta. Num procedimento não-segmentado, as informações de todas as camadas do calorímetro são agrupadas num único sinal sobre o qual são aplicadas as rotinas de processamento. O processo de extração de características por ICA é precedido por uma compactação através de PCA. A seguir serão descritos os fundamentos dessas técnicas estatísticas de processamento de sinais.

2.1 Análise de Componentes Independentes

Considerando que o sinal N-dimensional observado $\mathbf{x}$ é formado por uma combinação linear das fontes $\mathbf{s}$, a análise de componentes independentes (ICA - *Independent Component Analysis*) tem como objetivo encontrar o vetor de fontes a partir da observação apenas do vetor de misturas. Na forma matricial pode-se escrever (Hyvarinen et al., 2001):

$$\mathbf{x} = \mathbf{A}\mathbf{s} \tag{1}$$

onde $\mathbf{A}$ é a matriz de mistura.

Se os fatores a_{ij} são conhecidos o problema é facilmente resolvido a partir de:

$$\mathbf{s} = \mathbf{B}\mathbf{x} \tag{2}$$

onde $\mathbf{B} = \mathbf{A}^{-1}$. Na pratica, tanto as fontes s_i como os a_{ij} devem ser obtidos apenas dos sinais misturados x_i. No modelo básico da ICA (equação 1) assume-se que a matriz $\mathbf{A}$ é quadrada e não são considerados atrasos temporais nem a existência de ruído aditivo.

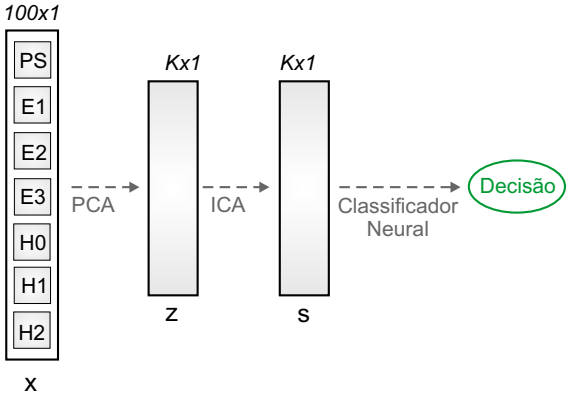

Figura 6: Diagrama da ICA sobre os sinais em soma de anéis.

Uma solução para o problema exposto na equação 1 pode ser obtida assumindo que as fontes s_i são estatisticamente independentes, o que é uma consideração realista na maioria dos casos e que não precisa ser exatamente verdadeira na prática. Existem muitas formas matemáticas para calcular os coeficientes b_{ij}. A descorrelação não-linear e a maximização da não gaussianidade são as mais utilizados (Hyvarinen and Oja, 2000).

Neste trabalho o algoritmo FastICA foi utilizado para estimar as componentes independentes. A medida da gaussianidade escolhida foi a negentropia, calculada por cumulantes de ordem superior e expansões polinomiais. Um método de ponto fixo é aplicado para calcular iterativamente a matriz B (Hyvarinen et al., 2001):

$$\mathbf{B} \longleftarrow E\{\mathbf{x}\mathbf{g}(\mathbf{B}^{\mathbf{T}}\mathbf{x}) - E\{\mathbf{g}^{'}(\mathbf{B}^{\mathbf{T}}\mathbf{x})\}\} \tag{3}$$

onde $E\{.\}$ é o operador esperança e $g(y)$ uma função não-linear geralmente escolhida entre:

$$g_1(y) = \tanh(y) \tag{4}$$
$$g_2(y) = y\exp(-y^2/2) \tag{5}$$
$$g_3(y) = y^3 \tag{6}$$

2.2 Compactação por PCA

A análise de componentes principais (PCA - *Principal Component Analysis*) é uma técnica estatística de processamento de sinais também conhecida como transformação de *Karhunen-Loève* (Haykin, 2001). O objetivo da PCA é encontrar uma representação do sinal de entrada onde grande parcela da energia esteja concentrada num pequeno número de componentes.

Usando informações da estatística de segunda ordem a transformação por PCA busca um espaço onde as projeções das entradas sejam não-correlacionadas e tenham máxima variância.

Considerando-se um vetor $\mathbf{x}$ aleatório de média nula com N elementos, a projeção y_i de $\mathbf{x}$ na direção de $\mathbf{w}_i$ pode ser expressa por:

$$y_i = \mathbf{w}_i^T \mathbf{x} = \sum_{k=1}^{N} w_{ki} x_k \tag{7}$$

Seguindo a definição da PCA, y_1 tem máxima variância, logo $\mathbf{w}_1$ é encontrado pela maximização de (Hyvarinen et al., 2001):

$$J_1^{PCA}(\mathbf{w}_1) = \mathcal{E}\{y_i^2\} = \mathbf{w}_1^T \mathbf{C}_x \mathbf{w}_1 \tag{8}$$

onde $\mathbf{C}_x$ é a matriz de covariância de $\mathbf{x}$.

A solução para o problema de maximização da equação 8 pode ser encontrado na algebra linear em função dos autovetores $\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_N$ da matriz $\mathbf{C}_x$. A ordem dos autovetores é tal que os autovalores associados satisfazem $d_1 > d_2 > ... > d_N$.

3 Teste de hipóteses

Após a extração de características por ICA sobre os sinais compactados através de PCA, o teste de hipóteses foi realizado utilizando classificadores neurais (ver Figura 6). As redes neurais, quando treinadas como classificadores, são capazes de estabelecer superfícies não-lineares de separação entre as classes.

Redes tipo perceptron de duas camadas (Haykin, 2001) foram treinadas, de modo supervisionado, para maximizar a eficiência de discriminação das classes. Para o ajuste dos pesos da rede foi utilizado o algoritmo RPROP (*Resilient back-propagation*) (Riedmiller and Braun, 1993). Os classificadores neurais utilizados possuem uma camada oculta com neurônios tipo tangente hiperbólica. A quantidade de neurônios na camada oculta variou com o número de componentes obtidos após a compactação por PCA, e foi escolhida visando a maximização do desempenho de discriminação das classes.

Para a escolha da dimensão da camada oculta, o número de neurônios foi gradativamente aumentado, realizando-se, a cada incremento, uma série de inicializações. Foi escolhida, então, a topologia de melhor desempenho. No caso de desempenho semelhante a menor rede foi utilizada.

4 Resultados

O conjunto de eventos disponível, obtido através de simulações de Monte Carlo para colisões próton-próton (ATLAS-Collaboration, 1999), foi dividido em dois grupos, um utilizado para o projeto (treino), e outro para o teste de desempenho das técnicas propostas. Os 2 conjuntos são compostos por assinaturas de: 11283 elétrons e 3774 jatos (treino), e 11298 elétrons e 3735 jatos (teste). O tipo de partícula correspondente a cada sinal é certificado pelo simulador. A interação com os subsistemas do detector e os efeitos da filtragem de primeiro nível são considerados durante as simulações, justificando a diferença no número de elementos das duas classes, uma vez que o LVL1 já eliminou grande parte dos jatos.

Como figuras de mérito do desempenho de classificação, foram utilizadas nesse trabalho a curva ROC (*Receiver Operating Characteristics*), obtida calculando a probabilidade de detecção (PD) e a probabilidade de falso alarme (PF) variando-se o patamar de decisão (Trees, 2001) e o índice SP, definido na equação 9:

$$SP = (Ef_1 + Ef_2)(Ef_1 \times Ef_2) \tag{9}$$

onde Ef_1 e Ef_2 são respectivamente a eficiência de detecção para as classes C_1 e C_2. O patamar escolhido é aquele que maximiza o SP.

Utilizando a abordagem proposta, os sinais dos calorímetros foram inicialmente compactados por PCA (para diferentes níveis de retenção de energia) e em seguida realizada a ICA. A extração das componentes independentes é capaz de revelar características do perfil de deposição de energia das partículas que estavam ocultas nos sinais originais, facilitando o processo de classificação. Na Tabela 1 são mostrados os índices SP calculados para cada um dos níveis de compactação. Percebeu-se que o desempenho de discriminação melhora com o aumento do nível de compactação por PCA, até 70% de retenção de energia. A partir daí, quanto mais componentes são desprezadas pior a performance do método.

Esse resultado é justificado pois a PCA não tem compromisso com a classificação, mas sim com a reconstrução dos sinais. No processo de compactação, componentes de baixa energia mas com alto poder de discriminação podem eventualmente ser eliminadas.

Uma forma de compactação que leva em conta o desempenho de classificação é a análise da relevância. A relevância R_i da componente independente IC_i é definida como o erro quadrático médio na saída do classificador quando a IC_i é

Tabela 1: Número de componentes e SP para os diferentes níveis de compactação por PCA.

Energia preservada	Número de Componentes	Produto SP
100%	100	1,80
99%	88	1,81
95%	58	1,82
90%	41	1,82
85%	31	1,82
80%	26	1,83
70%	17	1,83
60%	11	1,81

substituída por sua média no conjunto de treino (Gruber et al., 1994):

$$R_i = \frac{1}{N}\sum_{k=1}^{N}[Y_k - Y_{k,i(X_{k,i}=\overline{X_{k,i}})}]^2 \qquad (10)$$

onde N é o número de elementos do conjunto de treino.

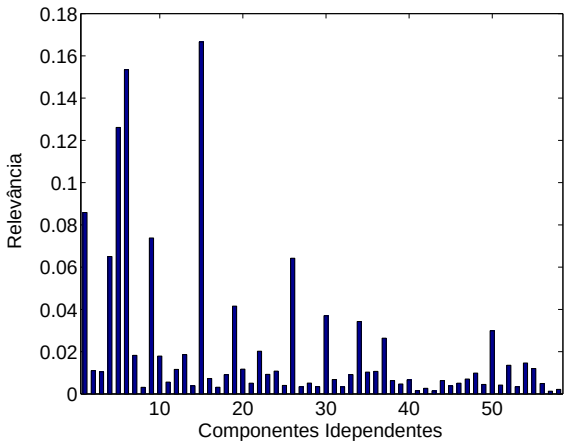

Figura 7: Relevância das componentes independentes para 95% de compactação por PCA.

A análise de relevância foi aplicada aos sinais compactados por PCA com retenção de 95% da energia. Obteve-se melhora no desempenho de classificação com a eliminação de até três das componentes menos relevantes, a supressão de quatro ou mais componentes diminuiu a eficiência de discriminação. ICs com alta relevância apresentam distribuição das classes com pouca sobreposição (ver Figura 8) facilitando o processo de discriminação. Percebeu-se que nessas componentes, a distribuição dos elétrons é sempre mais concentrada que a dos jatos (característica também presente no perfil de deposição dessas partículas). Componentes com baixa relevância apresentam maior sobreposição das distribuições de probabilidade (ver Figura 9), não contribuindo de forma significativa para o processo de separação. A Figura 10 mostra as curvas ROC e o produto SP para diferentes níveis de retenção de energia por PCA e compactação por relevância.

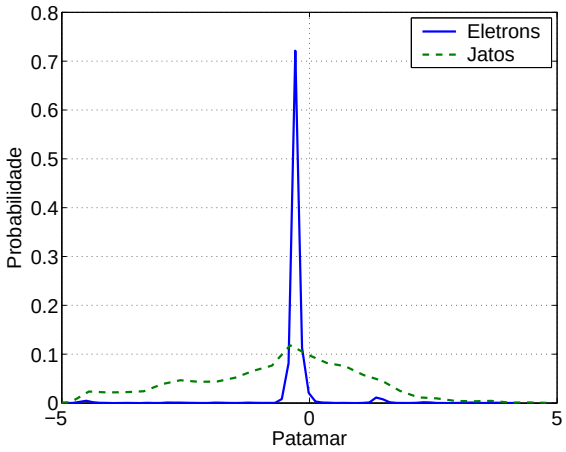

Figura 8: Distribuição de probabilidade da componente independente dezeseis para elétrons (linha contínua) e jatos (linha tracejada).

Na Tabela 2 pode-se observar as probabilidades de detecção e de falso alarme para os métodos de extração de característica propostos, destacando o melhor desempenho quando comparados ao T2Calo.

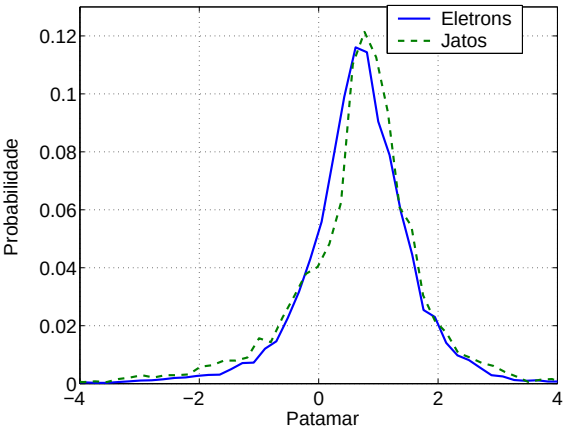

Figura 9: Distribuição de probabilidade da componente independente quarenta e um para elétrons (linha contínua) e jatos (linha tracejada).

No trabalho (Anjos et al., 2006), os tempos de processamento do T2Calo e de um classificador neural operando sobre os sinais formatados em anéis são comparados, sendo que a última estratégia é ligeiramente mais rápida. Considerando que, na técnica proposta nesse, apenas duas multiplicações matriciais (referentes a PCA e ICA) são adicionadas, não se espera significativo aumento no custo computacional, o que torna o método proposto compatível com os requisitos de latência do HLT. Testes criteriosos de desempenho precisam ser realizados com a técnica proposta, implementada no ambiente real de processamento.

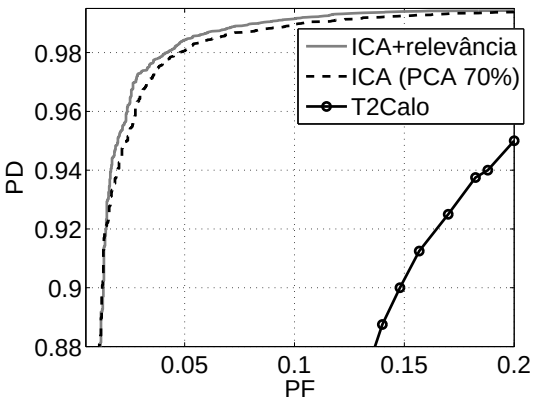

Figura 10: Curvas ROC obtidas para as técnicas propostas e para o T2Calo.

Tabela 2: Comparação das probabilidade de detecção (PD) e de falso alarme (PF).

Extração de características	**SP**	**PD** (%)	**PF** (%)
ICA + Relevância	**1,84**	99	10
ICA+PCA 70%	**1,83**	98	10
T2Calo	**1,60**	76	10

5 Conclusões

Uma nova estratégia foi proposta para a extração de características no canal elétron/jato no segundo nível de filtragem do detector ATLAS. Foi aplicada ICA sobre sinais compactados por PCA e resultados promissores foram obtidos usando um classificador neural no teste de hipóteses. Após análise da relevância de classificação, a rede foi novamente treinada e chegou-se a 97% de eficiência de discriminação de elétrons para 2,48% de probabilidade de falso alarme. Os resultados obtidos superam o do algoritmo atualmente adotado pela colaboração do ATLAS.

O modelo da ICA obteve bons resultados na representação dos sinais formatados em anéis. Porém, para preservar a estrutura em camadas dos calorímetros, pode-se utilizar um processamento segmentado (Simas et al., 2007), onde as informações de cada camada são tratadas separadamente. Na prática, sabe-se que os sinais dos calorímetros são modificados por fenômenos não-lineares, assim, um modelo de análise de componentes independentes não-linear (Hyvarinen et al., 2001) pode ser mais adequado. Estudos considerando as modificações expostas estão sendo conduzidos.

Agradecimentos

Os autores agradecem à CAPES, CNPQ e FAPERJ pelo apoio financeiro e aos colegas da colaboração Trigger/DAQ pelas contribuições e pelo fornecimento dos dados utilizados.

Referências

Anjos, A., Torres, R., Seixas, J., Ferreira, B. and Xavier, T. (2006). Neural triggering system operating on high resolution calorimetry information, *Nuclear Instruments and Methods in Physics Research* (559): 134–138.

ATLAS-Collaboration (1999). Atlas detector and physics performance technical design report, volume 1, *Technical report*, CERN.

ATLAS-Collaboration (2003). Atlas high-level trigger, data acquisition and controls technical design report, *Technical report*, CERN.

Gruber, A., Möck, J., Fröchtenicht, J. and Goldner, D. (1994). Performance of backpropagation networks in the second level trigger of the h1-experiment, *New Computing Techniques in Physics Research* .

Haykin, S. (2001). *Redes Neurais, Princípios e Prática*, Bookman.

Hyvarinen, A., Karhunen, J. and Oja, E. (2001). *Independent Component Analysis*, John Wiley and Sons.

Hyvarinen, A. and Oja, E. (2000). Independent component analysis: Algorithms and applications, *Neural Networks* (13): 411–430.

Martin, B. R. (2006). *Nuclear and Particle Physics*, John Wiley.

Price, M. (2002). The lhc project, *Nuclear Instruments and Methods in Physics Research* pp. 46–61.

Riedmiller, M. and Braun, H. (1993). A direct adaptive method for faster backpropagation learning, the rprop algorithm, *Proceedings of the IEEE International Conference on Neural Networks* .

Simas, E., Caloba, L. and Seixas, J. (2007). Segmented independent component analysis for online filtering using highly segmented detectors, *7th International Conference on Intelligent Systems Design and Applications - ISDA07* .

Trees, H. L. V. (2001). *Detection, Estimation, and Modulation Theory, Part I*, John Wiley and Sons.