Curvas Principais para a Seleção de Dados de Treinamento Neural com Grandes Volumes de Dados

Fernando Elias de Melo Borges, Danton Diego Ferreira
Departamento de Automática
Universidade Federal de Lavras
Lavras, Brasil
fborges@estudante.ufla.br, danton@ufla.br

José Manoel de Seixas
Laboratório de Processamento de Sinais
Universidade Federal do Rio de Janeiro
Rio de Janeiro, Brasil
seixas@lps.ufrj.br

***Abstract*—In environments with big data problems, to make a smart data selection with the goal of training machine can be essential to reduce the computational demand of the application. This paper presents a method based on principal curves for data selection during the neural networks training in an experiment of particle collision with high events rate. The method used real data of collision and it accomplished 3 selection approaches through mapping of Euclidean distances in each event to the respective Principal Curve. Preliminary results in the classification of neural networks presented low differences using the selection method and considerable reduction in the training time.**

Keywords— Principal Curves; Data Selection; Big Data.

I. INTRODUÇÃO

Aplicações com *big data* (grandes volumes de dados) e modelos de aprendizagem de máquina cada vez mais complexos vêm sendo usados com mais frequência atualmente e têm mostrado capacidade de obtenção de resultados promissores em diversas aplicações. Alguns exemplos são a identificação de padrões de mobilidade urbana, por meio de dados obtidos de telefones celulares [1]; análises de condições climáticas [2], dentre outras. Contudo, tais modelos requerem grandes ciclos de desenvolvimento, necessitando de grande poder computacional e demandando muito tempo para processar tal massa de dados.

Na área de física de altas energias também são encontrados problemas envolvendo *big data*. Tais problemas são ocorridos devido aos eventos de interesse serem de rara ocorrência, o que leva aos sistemas de aquisição de dados terem uma alta taxa de eventos (em torno de 70TB/s) para ser obtida a estatística necessária para as análises. Além disto, o sistema de calorimetria utilizado na aquisição de dados possui grande número de canais e fina granularidade, o que leva a geração de dados com elevada dimensionalidade e grande volume.

Dentro da aplicação mencionada, o experimento ATLAS inserido no Centro Europeu para a Pesquisa Nuclear (CERN) tem em seus trabalhos problemas envolvendo *big data* para estudar os eventos físicos de interesse ocorridos no interior do colisor LHC (*Large Hadron Collider*). As colisões são realizadas por meio de cruzamento de prótons no interior do LHC. Estas colisões geram partículas instáveis que são analisadas por meio de seus decaimentos em partículas mais estáveis, como elétrons, por exemplo. No ATLAS, elétrons são objeto de estudo para a reconstrução dos eventos físicos de interesse ocorridos no colisor, como, por exemplo, a observação do Bóson de Higgs ocorrida em 2012 [3] [4].

Tais eventos, como a observação do Bóson de Higgs, são raros, o que justifica a necessidade da geração de grandes volumes de dados para aquisição da estatística necessária para a reconstrução dos eventos físicos de interesse. A maioria de dados gerados nestes eventos é composta por ruído de fundo do experimento, sem interesse de estudo.

Para evitar que eventos sem interesse de utilização sejam armazenados, algoritmos de filtragem *online* vêm sido desenvolvidos e aperfeiçoados ao longo do tempo para filtrar os sinais provenientes dos elétrons em meio a grande massa de ruído de fundo, armazenando, assim, somente os sinais de interesse. Dentre tais algoritmos, o *NeuralRinger*, [5] desenvolvido pela Coppe/UFRJ, extrai a informação do sistema de calorimetria, compactando-a na forma de anéis concêntricos de energia e utiliza um *ensemble* de redes neurais para realizar a função da filtragem *online*. O *NeuralRinger* possui como principais características a elevada taxa de eficiência e o baixo falso alarme, fazendo com que, além de detectar os sinais dos elétrons, evite a coleta de sinais de ruído de fundo erroneamente classificados como sinais de elétrons. Todavia, dada a problemática de *big data* envolvendo o ATLAS, realizar o ciclo de desenvolvimento de tais modelos neurais pode demandar grande esforço computacional para que ele atinja a convergência.

Baseando-se no problema supracitado, propor a realização de uma seleção inteligente de dados antes do treinamento do algoritmo, torna-se uma abordagem viável para reduzir a carga computacional do algoritmo de detecção desde que um desempenho similar ao treinamento com todo o conjunto de dados seja obtido. Tal prática pode ser vista em [6] em [7], em que foi apresentada a problemática para seleção de parâmetros em *big data* e algumas soluções foram propostas.

Neste trabalho é proposta uma alternativa de seleção de dados utilizando Curvas Principais (CP) [8]. As CP possuem algoritmos eficientes para sua extração, como o k-segmentos [9], que foi utilizado neste trabalho. As CP têm a capacidade de gerar representações compactas dos dados, explorando

correlações não-lineares dos mesmos. A partir da seleção de dados, foram testadas abordagens de seleção em uma rede neural do tipo *perceptron* multicamadas (MLP), simulando o processo realizado pelo algoritmo *NeuralRinger* utilizando todo o conjunto de dados utilizado no desenvolvimento da CP e conjuntos de dados reduzido, explorando o mapeamento do conjunto de dados utilizando as curvas obtidas.

II. MÉTODO PROPOSTO

A. Curvas Principais

As Curvas Principais representam uma generalização não linear da Análise de Componentes Principais (PCA) [8] e têm sua forma sugerida pelo conjunto de dados que é usado para construí-las. As CP extraem um modelo unidimensional de um conjunto de dados multidimensional, resultando numa compactação do mesmo. Há alguns algoritmos de extração de CP na literatura. Dentre eles, destaca-se o k-segmentos não suave [9]. Este algoritmo constrói as CP de maneira incremental, por segmentos de reta, e possui menor tendência a mínimos locais e convergência prática garantida. O algoritmo requer 3 passos para execução, conforme ilustrado no diagrama em blocos da Fig. 1, sendo:

Passo 0: consiste na inserção do primeiro segmento, usando-se todo o conjunto de dados, e tomando a direção da primeira componente principal com um tamanho 3/2 do desvio padrão desta componente;

Passo 1: inserção do segundo segmento e redefinição no ponto central do agrupamento. Os eventos pertencentes ao agrupamento são definidos pelo algoritmo *k-means* baseado nas regiões de Voronoi, estas alocam os eventos mais próximos do centro da região que dos segmentos da curva. A união dos segmentos se dá por uma linha reta os segmentos nos conjuntos onde houver mudança têm seus tamanhos recalculados;

Passo 2: consiste no teste de convergência do algoritmo, podendo ser de duas diferentes formas. A primeira testa o número k de segmentos, verificando se corresponde ao número máximo de segmentos pré-estabelecidos pelo usuário K_{max} ($k = K_{max}$). A segunda verifica se o menor agrupamento obtido possui, ao menos, 3 segmentos. Não satisfeitas tais condições, o algoritmo retorna ao passo 1.

B. Método Proposto

O método proposto é baseado no uso das CP para a seleção de dados para o treinamento de uma rede neural com o objetivo de detecção de elétrons. Foram utilizados, para o desenvolvimento das CP, dados reais de colisão do tipo $Z \rightarrow ee$ do ano de 2018. Estes dados eram compostos por padrões de sinal de elétrons e por ruído de fundo.

Como conjunto de desenvolvimento, foram utilizados 50.000 eventos que consistem na informação anelada dos sinais provenientes do sistema de calorimetria do ATLAS realizado pelo *NeuralRinger* [5] que compacta os sinais do detector na forma de anéis de energia. Tais dados foram selecionados aleatoriamente da base de dados de colisão. Em seguida, utilizando o conjunto selecionado para o desenvolvimento, foram geradas duas CP, sendo uma para o padrão de sinal e uma para o padrão de ruído de fundo. Extraídas as curvas para cada padrão, foi realizado um mapeamento das distâncias euclidianas de cada evento à sua respectiva CP, ranqueando-as a partir dos valores de distância obtidos. Por fim, a seleção de dados é feita por meio deste ranqueamento. Para tal, foram propostas 3 abordagens de seleção:

1. Utilizando apenas os dados mais pertos da curva. Variando o tamanho do conjunto reduzido com os seguintes tamanhos de conjunto: 1.000, 2.000, 5.000 e 10.000 eventos;
2. Utilizando apenas os dados mais distantes à curva. Variando o tamanho do conjunto reduzido com os seguintes tamanhos de conjunto: 1.000, 2.000, 5.000 e 10.000 eventos;
3. Utilizando tamanho do conjunto reduzido fixo, com dados mais próximos e mais distantes à curva e variando a proporção dos dados utilizados (mais próximos e mais distantes à curva). O tamanho do conjunto de dados foi fixado em 10.000 eventos.

Tais abordagens visam uma análise de verificação de qual conjunto de dados, de acordo com seu agrupamento pelo ranqueamento das distâncias, possui melhor representatividade na redução do conjunto (abordagens 1 e 2), ou se uma mistura ponderada dos dados se faz mais eficiente dado ao agrupamento alocar os dados com maior distância à CP (abordagem 3).

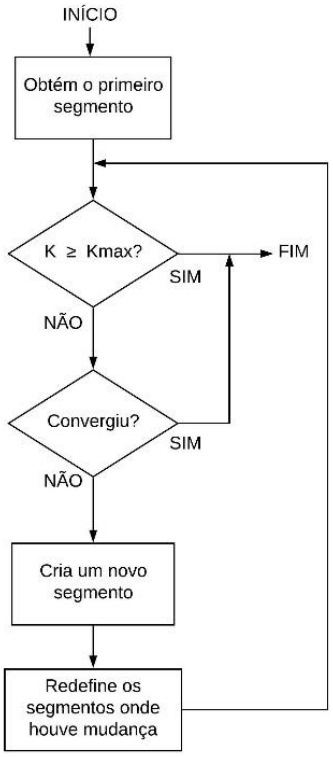

Fig. 1. Fluxograma do algoritmo k-segmentos para obtenção de CP.

Após a seleção dos dados, foi realizado o projeto do classificador, baseando-se no algoritmo *NeuralRinger* [5], para isto, foi utilizada uma rede neural do tipo *perceptron* multicamadas (MLP – *Multi-Layer Perceptron*) com topologia de 10 neurônios na camada escondida e função de ativação tangente hiperbólica. O ajuste de parâmetros se deu pelo método de otimização Quasi Newton. O processo de treinamento utilizou validação cruzada do tipo *k-fold* com número de *folds* (partições) igual à 10, ou seja, variando o conjunto de treino e validação durante 10 ciclos de execução. Em cada ciclo de execução foram tomadas as medidas de desempenho como: a probabilidade de detecção (P_D), a probabilidade do falso-alarme que corresponde aos padrões de ruído de fundo erroneamente classificados como sinais de elétron (P_F) e o índice soma-produto (SP), este último dado pela equação (1).

$$SP = \sqrt{\sqrt{P_D(1-P_F)}\,\frac{P_D+(1-P_F)}{2}} \qquad (1)$$

Após o processo de desenvolvimento da rede neural, foi escolhido o modelo selecionado para o teste. Tal escolha se deu pelo maior valor de SP (SP_{max}) obtido durante o processo de validação cruzada no treinamento. Os tamanhos dos conjuntos de dados utilizados no desenvolvimento da CP e de teste das redes estão descritos na Tabela I. Para efeito comparativo, o conjunto de teste foi o mesmo para todas as abordagens de seleção.

TABELA I. CONJUNTO DE DADOS UTILIZADO

Classe	Conjunto de desenvolvimento	Conjunto de teste
Sinal	50.000	2.950.000
Ruído de Fundo	50.000	150.000

III. RESULTADOS E DISCUSSÃO

Foi extraída uma CP para cada padrão (sinal e ruído de fundo), as CP obtidas foram formadas por 32 segmentos cada. Após a extração de cada CP, as distâncias de cada evento à sua respectiva curva foram calculadas. A Fig. 2 mostra um gráfico das distâncias medidas enquanto um histograma das distâncias é apresentado na Fig. 3.

Em seguida, foi realizado o projeto das redes neurais de acordo com cada abordagem de redução. Os resultados de P_D, P_F e SP para o treinamento e os resultados de testes com os dados nas 3 abordagens de seleção seguem, respectivamente nas Tabelas II, III, IV. A Tabela II mostra os resultados para a abordagem 1 (utilizando os dados mais próximos à curva) em relação ao tamanho do conjunto de dados selecionado (N_{t1}), a Tabela III para a abordagem 2 (dados mais distantes à curva) em relação ao tamanho do conjunto de dados selecionado (N_{t2}) e a Tabela IV apresenta os resultados utilizando os dados mais próximos junto com os dados mais distantes de acordo com o percentual dos dados mais próximos à curva (pp). Os resultados utilizando todo o conjunto de dados de desenvolvimento da CP (N_t) são descritos na Tabela V.

Após análise dos resultados gerados pelas redes neurais contidos nas Tabelas II, III e IV e comparando com os resultados de referência da Tabela V, podem-se destacar alguns pontos: (i) Para a abordagem 1, houve resultados de falso alarme melhores que nas demais abordagens, contudo, os resultados de detecção se mostraram mais baixos dentre as mesmas. (ii) Os resultados de detecção para a abordagem 2 se mostraram os maiores dentre as 3 abordagens, entretanto, os resultados de falso alarme foram elevados tanto em valor quanto em desvio-padrão. (iii) Para a abordagem 3, houve valores medianos com relação aos resultados das abordagens de seleção 1 e 2, porém, interessantes para análise devido à baixa queda dos valores de detecção e a valores de falso alarme com desvio-padrão menor que na abordagem 2. (iv) Os valores obtidos nos resultados podem chegar à valores de resultados com proximidade à referência ou, em alguns casos, com resultados de detecção (resultados das abordagens 2 e 3) e falso alarme (resultados da abordagem 1) melhores que os resultados usando todo o conjunto de dados.

Observando os resultados de classificação obtidos e realizando uma análise conjunta com os resultados das distâncias nas Fig. 2 e Fig. 3, pode-se observar que a maioria dos dados está nas baixas distâncias e que a discrepância das distâncias altas e baixas é, relativamente, alta. Tal situação pode afetar os resultados de agrupamento que pode influenciar diretamente nos resultados de classificação, como, por exemplo, o desvio-padrão dos resultados. Para isto, uma abordagem baseada em análise de agrupamentos por meio dos segmentos da CP pode ser interessante, realizando uma outra forma de seleção de dados e, possivelmente, reduzindo as discrepâncias no mapeamento.

Visando os melhores resultados de teste, os resultados que obtiveram, em conjunto, os melhores resultados de detecção e falso alarme foram os resultados utilizando a abordagem 3 (dados mais próximos em conjunto com dados mais distantes) com percentual de 80% ou 70% de dados mais próximos à curva. Tais conjuntos obtiveram valores de detecção maiores que a referência (tanto a média quanto o desvio-padrão) e valor de falso alarme pouco maior que os valores de referência apesar do desvio ser maior, para uso de um conjunto de dados de treinamento maior, o problema do desvio-padrão poderá ser corrigido.

Outro aspecto importante a se considerar é o tempo de execução do algoritmo, em que, utilizando o conjunto de dados total (com 50.000 eventos) o tempo de execução foi em torno de 23 minutos. Com o uso de conjuntos de dados reduzido, foram obtidos tempos de 5 minutos (utilizando 10.000 eventos, tomando-se o maior dos tempos de execução). Logo, o método proposto consegue atingir resultados similares entre os conjuntos de dados com redução significativa no tempo de execução da rede neural, neste caso, a redução chegou em mais de 75% do tempo de execução, em comparativo com o tempo gasto no treinamento da rede neural com o conjunto com todos os dados.

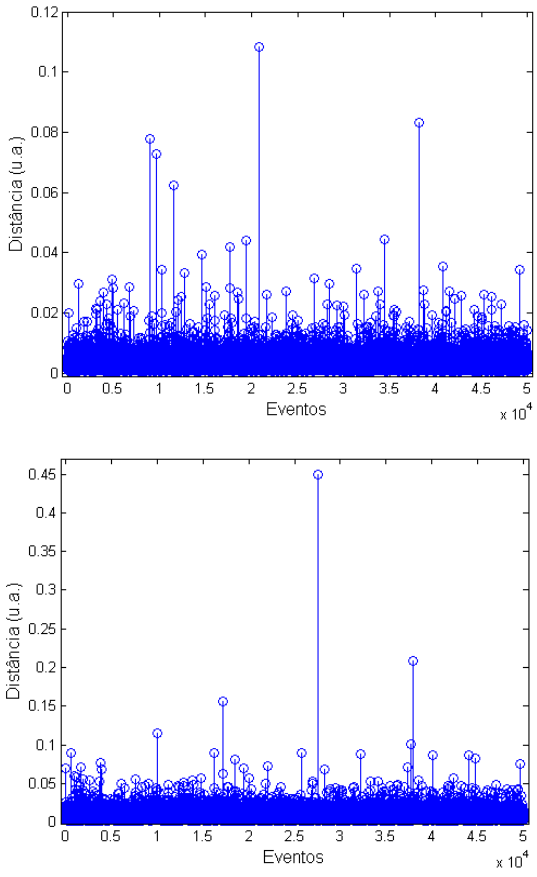

Fig. 2. Distâncias de cada evento à sua respectiva CP. (a) sinal e (b) ruído de fundo.

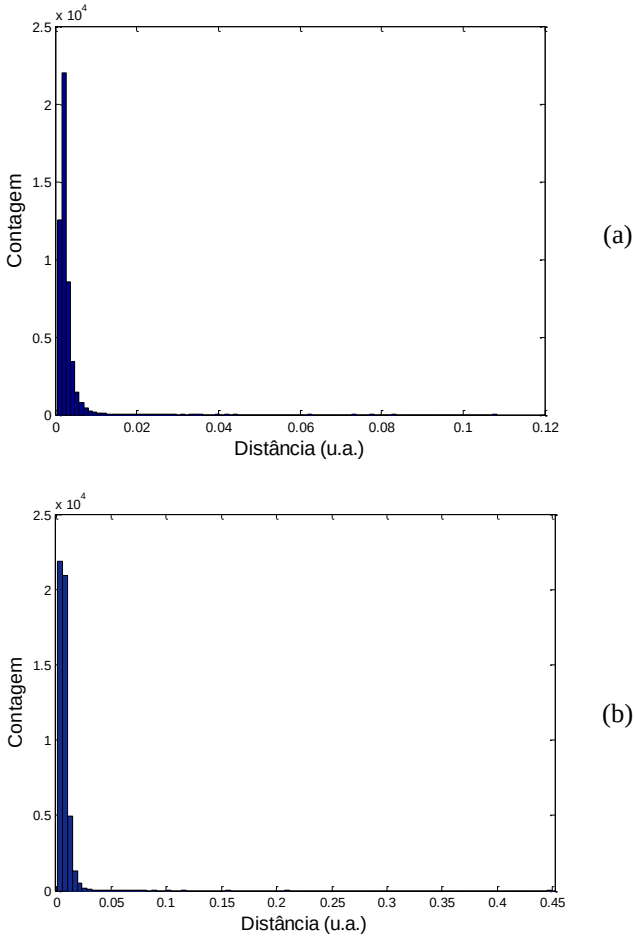

Fig. 3. Histograma das distâncias dos eventos à sua respectiva CP. (a) sinal e (b) ruído de fundo.

IV. CONCLUSÕES

Por meio de testes, utilizando um conjunto de dados experimental contendo dados reais de colisão de 2018, e analisando os resultados preliminares pode-se observar um potencial do método de seleção de dados utilizando Curvas Principais. Os resultados apresentados constaram valores de detecção próximos ou maiores que a referência (para as abordagens 2 e 3) e valores de falso alarme menores que a referência (caso da abordagem 1). Tais resultados propiciam o avanço da técnica de seleção para experimentos futuros visando uma análise mais aprofundada da técnica de seleção, formas de melhorias da mesma, visando uma redução dos valores de desvio-padrão dos resultados, por exemplo. Para os próximos passos, visa-se um estudo da seleção dos dados aplicados ao algoritmo *NeuralRinger* para comparativo dos resultados de treinamento do algoritmo com conjunto de dados reduzidos e seus efeitos; também serão realizadas análise de quadrante para verificação do comportamento do algoritmo com um novo tipo de conjunto de dados, além da análise de impacto visando analisar se o um conjunto de dados reduzido poderá introduzir ou não tendências nos resultados. Tais estudos com o objetivo de, assim, colaborar com o melhoramento do método de detecção de elétrons, visando manter os bons resultados reduzindo o custo computacional.

TABELA II. RESULTADOS PERCENTUAIS (MÉDIA±DESVIO-PADRÃO) PARA ABORDAGEM DE SELEÇÃO 1:

N_{t1}	Treino			Teste	
	SP_{max}	P_D	P_F	P_D	P_F
10.000	99,98 ±0,00	100,00 ±0,00	0,04 ±0,79	94,43 ±21,60	0,90 ±9,06
5.000	100,00 ±0,00	100,00 ±0,00	0,00 ±0,00	94,59 ±21,63	0,93 ±9,23
2.000	100,00 ±0,00	100,00 ±0,00	0,00 ±0,00	94,17 ±21,12	1,00 ±9,18
1.000	100,00 ±0,00	100,00 ±0,00	0,00 ±0,00	94,96 ±20,07	1,06 ±9,60

TABELA III. RESULTADOS PERCENTUAIS (MÉDIA±DESVIO-PADRÃO) PARA ABORDAGEM DE SELEÇÃO 2:

N_{t2}	Treino			Teste	
	SP_{max}	P_D	P_F	P_D	P_F
10.000	97,21 ±0,00	97,1 ±7,03	2,69 ±14,03	98,67 ±3,93	3,76 ±16,41
5.000	97,26 ±0,00	97,14 ±10,41	2,61 ±14,83	97,95 ±3,3	5,17 ±20,26
2.000	97,31 ±0,00	96,31 ±16,29	1,69 ±12,39	99,49 ±3,21	7,6 ±25,57
1.000	99,18 ±0,00	99,54 ±0,71	1,18 ±10,03	99,24 ±2,21	12,24 ±32,08

TABELA IV. RESULTADOS PERCENTUAIS (MÉDIA±DESVIO-PADRÃO) PARA ABORDAGEM DE SELEÇÃO 3:

pp [%]	Treino			Teste	
	SP_{max}	P_D	P_F	P_D	P_F
90	99,27 ±0,00	99,31 ±6,87	0,76 ±7,07	98,69 ±9,06	2,37 ±13,47
80	98,70 ±0,00	98,97 ±7,31	1,57 ±10,01	99,00 ±6,27	2,60 ±13,68
70	98,24 ±0,00	97,80 ±10,58	1,32 ±8,68	99,03 ±5,34	2,76 ±13,9
60	98,18 ±0,00	98,07 ±9,15	1,70 ±10,95	99,06 ±4,64	2,81 ±14,17
50	97,72 ±0,00	97,41 ±8,64	1,96 ±11,43	99,09 ±4,44	3,00 ±14,5
40	97,81 ±0,00	97,97 ±7,45	2,36 ±13,01	98,95 ±4,40	2,97 ±14,63
30	97,79 ±0,00	97,39 ±9,11	1,80 ±10,52	98,83 ±4,35	2,98 ±14,74
20	97,40 ±0,00	97,25 ±9,39	2,45 ±13,38	98,74 ±4,17	3,21 ±15,09
10	97,09 ±0,00	97,05 ±9,59	2,87 ±15,25	98,40 ±4,35	3,31 ±15,4

TABELA V. RESULTADOS PERCENTUAIS (MÉDIA±DESVIO-PADRÃO) PARA TODO O CONJUNTO DE DADOS:

Nt	Treino			Teste	
	SP_{max}	P_D	P_F	P_D	P_F
50.000	98,29 ±0,00	98,2 ±7,73	1,61 ±10,21	98,15 ±7,62	1,79 ±10,6

AGRADECIMENTOS

Agradecimentos à CAPES, CNPq, FAPERJ e FAPEMIG pelo apoio a este trabalho.

REFERÊNCIAS

[1] Jiang, S.; Ferreira, J.; Gonzalez, M. C. Activity-based human mobility patterns inferred from mobile phone data: A case study of Singapore. IEEE Transactions on Big Data, v. 3, n. 2, p. 208-219, 2017.

[2] Onal, A. C. et al. Weather data analysis and sensor fault detection using an extended iot framework with semantics, big data, and machine learning. In: 2017 IEEE International Conference on Big Data (Big Data). IEEE, 2017. p. 2037-2046.

[3] ATLAS COLLABORATION. Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC", Phys. Lett. B, v. 716, pp. 1, 2012. doi: 10.1016/j.physletb.2012.08.020.

[4] CMS COLLABORATION. Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC", Physics Letters B, v. 716, n. 1, pp. 30-61, 2012. ISSN: 0370-2693. doi: http://dx.doi.org/10.1016/j.physletb.2012.08.021.

[5] Freund, W.S. IDENTIFICAÇÃO DE ELÉETRONS BASEADA EM UM CALORÍMETRO DE ALTAS ENERGIAS FINAMENTE SEGMENTADO. 2018. Tese de Doutorado. Universidade Federal do Rio de Janeiro.

[6] Li, J.; Liu, H. Challenges of feature selection for big data analytics. IEEE Intelligent Systems, v. 32, n. 2, p. 9-15, 2017.

[7] Rong, M.; Gong, D.; Gao, Xi. Feature Selection and Its Use in Big Data: Challenges, Methods, and Trends. IEEE Access, v. 7, p. 19709-19725, 2019.

[8] Hastie, T. J., Stuetzle, W., "Principal Curves", Journal of the American Statistical Association, v. 84, n. 406, pp. 502-516, 1989.

[9] Verbeek, J. J., Vlassis, N., Krose, B., "A soft k-segments algorithm for principal curves", Proceedings of International Conference on Artificial Neural Networks, pp. 450-456, 2001.