DOI: 10.48011/asba.v2i1.1359

Aprendizagem em grandes volumes de dados: Seleção de Dados para Treinamento de Máquina em Ambientes com Alta Taxa de Eventos

Fernando E. M. Borges * **Danton D. Ferreira** **
José M. Seixas ***

* *Departamento de Automática, Universidade Federal de Lavras, MG (e-mail: fborges@estudante.ufla.br).*
** *Departamento de Automática, Universidade Federal de Lavras, MG (e-mail: danton@ufla.br)*
*** *Laboratório de Processamento de Sinais, Universidade Federal do Rio de Janeiro, RJ (email: seixas@lps.ufrj.br)*

Abstract: Environments with high rates of events generating large volumes of data are increasingly present. Large data acquisition and analysis systems, requiring complex mining tools have been adopted in several areas. This systems have requirements during their development and operation, such as processing time and memory consumption. Methods for smart data selection, which allow reducing the development time of machines in big-data environments, become a viable approach to optimize data mining-based systems. In this paper, the smart selection of events based on Principal Curves is proposed. The method exploits the nonlinear correlations of the data. The event selection is made by mapping the distances of each event to the Principal Curve. To test the method, a real dataset was used for the electron trigger system of the ATLAS experiment at CERN (European Center for Nuclear Research). After selecting the data, they are tested in a multilayer Neural Network (MLP) to simulate the real process of the system. Preliminary results showed similar detection and false alarm results between the complete and reduced sets.

Resumo: Ambientes com alta taxa de eventos gerando grandes volumes de dados estão cada vez mais presentes. Sistemas de aquisição e análise de dados, exigindo ferramentas de mineração complexas vêm sendo adotados em diversas áreas. Tais sistemas possuem requisitos durante seu desenvolvimento e operação, como tempo de processamento e consumo de memória. Propor um método de seleção inteligente de dados, que permita reduzir o tempo de desenvolvimento de máquinas em ambientes de *big-data*, atingindo resultados similares de desempenho entre os conjuntos de treinamento total e reduzidos pela seleção, torna-se uma abordagem viável na otimização do desenvolvimento destes sistemas. Neste artigo é proposta uma seleção inteligente de eventos utilizando Curvas Principais, que explora correlações não lineares dos dados. A seleção se dá pelo mapeamento das distâncias de cada evento à Curva Principal. Para testar o método, utilizou-se uma base real de dados referente ao sistema de filtragem online de elétrons do experimento ATLAS do CERN (Centro Europeu para a Pesquisa Nuclear). Após a seleção dos dados, estes são testados em uma Rede Neural do tipo multicamadas (MLP) para simular o processo real do sistema. Resultados preliminares apresentaram resultados de detecção e falso alarme similares entre os conjuntos completo e reduzido.

Keywords: Data Selection; Principal Curves; Big-Data; Machine Learning; Neural Networks
Palavras-chaves: Seleção de dados; Curvas Principais; Big-Data; Aprendizagem de Máquina; Redes Neurais

1. INTRODUÇÃO

Atualmente, com o advento da era da informação, o uso de ferramentas de IoT (Internet das coisas, do inglês *Internet of Things*), sistemas envolvendo grandes volumes de dados com elevada dimensionalidade em alta taxa de eventos, ou seja, sistemas que fazem uso de *big-data*, estão cada vez mais presentes. Implementações de sistemas de cidades inteligentes (Al Nuaimi et al., 2015), identificação de padrões de mobilidade urbana (Jiang et al., 2017), setor de saúde (Kumar and Singh, 2018) e monitoramento de condições climáticas (Onal et al., 2017) são alguns exemplos de aplicações envolvendo *big-data* nas mais diversas áreas.

Contudo, devido à complexidade dos dados, alta velocidade de aquisição e requisitos de desempenho computacional, aplicações que envolvem *big-data* vêm apresentando grandes desafios. Trabalhos, como os reportados por Jin et al. (2015) e Fan et al. (2014), fazem uma revisão acerca do assunto e relatam a importância e os desafios do uso de técnicas de aprendizagem de máquina em *big-data*. Dentre tais desafios, encontra-se o desafio de custo computacional, onde são necessárias abordagens que consumam o mínimo possível de processamento e memória, além de atingirem os requisitos de predição desejados.

Baseando-se nesta problemática, propor métodos em que se busca reduzir o tamanho do problema (número elevados de eventos e da dimensão dos dados) para os algoritmos de aprendizagem de máquina a serem treinados mantendo o desempenho do modelo, tem grande importância e aplicabilidade prática. Rong et al. (2019) apresentam uma revisão sobre seleção de atributos em ambientes envolvendo *big-data* e mostra alguns modelos utilizados em diferentes casos. Outras aplicações envolvendo seleção de atributos para fins de redução de dimensionalidade do problema são vistos em Fong et al. (2015) e Hasanin et al. (2019).

Neste artigo é proposto um método de seleção de eventos para treinamento de modelos de aprendizagem de máquina. Como objeto de estudo, foi utilizado o sistema de filtragem *online* de elétrons do experimento ATLAS do CERN (Centro Europeu para a Pesquisa Nuclear). O experimento ATLAS é imerso em um problema de *big-data* dadas as características dos estudos realizados onde eventos são coletados em taxas elevadas (cerca de 70TB/s de dados são gerados) para obtenção da estatística necessária para as análises. O sistema de calorimetria utilizado na aquisição dos dados possui um grande número de canais e fina granularidade, o que implica em dados de elevada dimensionalidade.

Os eventos ocorridos no interior do LHC (*Large Hadron Collider*) são provenientes de colisões de prótons emitidos em sentidos opostos, estas colisões geram partículas instáveis que podem ser analisadas por meio de seus decaimentos em partículas mais estáveis como, por exemplo, elétrons. No ATLAS, elétrons são objeto de interesse de estudo para reconstrução de eventos físicos raros de interesse, como a observação do Bóson de Higgs (Aad et al., 2012), (Chatrchyan et al., 2012).

Dada a conjectura do problema, a grande maioria dos sinais que passam pelos sistemas de aquisição são provenientes de ruído de fundo do experimento, sem interesse de estudo. Para evitar que tais sinais sejam armazenados, sistemas de filtragem online *Trigger* foram estudados e implementados tanto em *hardware* quanto em *software* de maneira que sejam armazenados apenas os sinais de interesse de estudo do experimento. Dentre tais algoritmos, encontra-se o *NeuralRinger* (Freund, 2018), desenvolvido pela COPPE/UFRJ, que extrai a informação do sistema de calorimetria, compactando-a em anéis de energia concêntricos e realiza a filtragem por meio de um *ensemble* de Redes Neurais. O *NeuralRinger* possui como características uma alta taxa de detecção de elétrons aliada a um baixo falso alarme, sendo, além de eficiente na detecção de partículas de interesse, eficiente em evitar aquisição de sinais de ruído de fundo erroneamente classificados como sinais de elétrons, sendo também, atualmente, o algoritmo de referência na filtragem online de elétrons do ATLAS.

Por outro lado, dada toda a problemática envolvendo *big-data* no ATLAS, tais modelos neurais vêm requerendo grande tempo de processamento para realizarem seu ciclo de desenvolvimento, demandando grande esforço computacional para que possam convergir e atingir os resultados esperados. Visando contornar tal problema, a proposta deste trabalho consiste em reduzir os conjuntos de dados por meio de uma seleção inteligente de eventos para o treinamento das Redes Neurais, gerando menor tempo de processamento dos dados sem perder desempenho.

Para isto, este trabalho propõe um método de seleção inteligente de dados utilizando Curvas Principais (CP) (Hastie and Stuetzle, 1989), por meio do algoritmo k-segmentos (Verbeek et al., 2002). O método tem por objetivo realizar uma seleção de eventos, reduzindo o tamanho do conjunto de dados que irão treinar as Redes Neurais com a mesma dimensionalidade. Esta seleção se dá pelo mapeamento das distâncias euclidianas dos eventos que projetaram a Curva e os conjuntos de dados são reduzidos por meio deste. Após a seleção de dados, foram realizados testes em uma Rede Neural Multicamadas (MLP) para fins de simulação do algoritmo *NeuralRinger* tanto para o conjunto total de dados quanto para os conjuntos reduzidos de dados utilizando a seleção e seus resultados foram comparados para verificação do método.

2. CURVAS PRINCIPAIS

As CP consistem em uma técnica de processamento dos dados, sendo uma generalização não-linear da Análise de Componentes Principais (ou PCA, do inglês *Principal Component Analysis*) (Hastie and Stuetzle, 1989). As CP possuem a capacidade de extrair modelos compactos de dados explorando correlações não-lineares entre eles. O algoritmo inicialmente proposto por Hastie apresentou limitações no tocante à seu uso prático, como *bias* nas CP geradas e uma não garantia de convergência do modelo. Visando contornar tais problemas, métodos alternativos para a extração de CP foram desenvolvidos (Banfield and Raftery, 1992; Tibshirani, 1992; Delicado, 2001). Dentre os algoritmos propostos, encontra-se o k-segmentos, proposto por Verbeek et al. (2002). Tal método extrai CP de maneira incremental por segmentos de reta e possui como vantagens a baixa tendência a mínimos locais e a convergência prática garantida. O algoritmo obtém as CP em 3 passos, que são descritos a seguir e estão ilustrados no fluxograma da Figura 1:

Passo 0: Consiste na inserção do primeiro segmento, este tem a direção da primeira componente principal e comprimento equivalente a 3/2 do desvio padrão dos dados.

Passo 1: Neste passo é inserido o segundo segmento e redefinido o ponto central de agrupamento dos dados. O agrupamento dos eventos é realizado via algoritmo *k-means* baseado nas regiões de Voronoi, alocando os eventos mais próximos à região do que aos seus vizinhos. Os segmentos são unidos por meio de uma linha reta (não-suave). Os segmentos que houveram alteração nos seus agrupamentos tem seus tamanhos recalculados. Para os demais segmentos o procedimento é o mesmo.

Passo 2: Consiste na análise da convergência do algoritmo; esta é realizada de duas maneiras, primeiro se o número de segmentos (k) atingiu o número máximo estipulado pelo usuário (K_{max}), ou seja, se $k = K_{max}$, ou se o menor agrupamento possui, no mínimo, 3 segmentos. Caso nenhuma destas condições sejam satisfeitas, o algoritmo retorna ao passo 1.

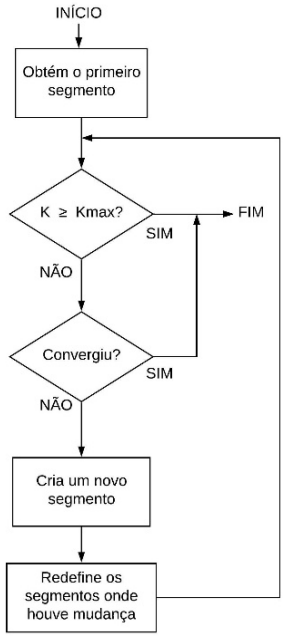

Figura 1. Fluxograma do algoritmo k-segmentos para obtenção de Curvas Principais.

3. MÉTODO APLICADO

Neste trabalho foi realizado como base de dados de estudo, dados do sistema de filtragem online de elétrons do experimento ATLAS do CERN. Os dados consistem em dados reais de colisão do ano de 2018 e correspondem à informação anelada de energia compactada pelo algoritmo anelador (Freund, 2018). As matrizes possuem dois padrões: sinal de elétrons e ruído de fundo do experimento (*background*). O tamanho do conjunto de dados utilizado neste experimento encontra-se na Tabela 1.

A partir do algoritmo k-segmentos realizado por Verbeek et al. (2002), foi realizada sua tradução para a linguagem Python, dada sua maior flexibilidade por se tratar do uso de uma ferramenta de código livre (*open-source*). Em posse dos dados obtidos, foram geradas uma CP para cada padrão do problema e explorado o mapeamento das distâncias euclidianas dos eventos à sua respectiva Curva. Por meio deste mapeamento, foram propostas 3 abordagens de seleção de dados para o treinamento das Redes Neurais:

Tabela 1. Tamanho do conjunto de dados utilizado

Padrão	Desenvolvimento	Teste
Sinal	200.000	3.300.000
Background	200.000	7.808

Abordagem 1: Foram selecionados os eventos mais próximos à CP, variando o tamanho do conjunto de dados reduzido (N_t);

Abordagem 2: Foram selecionados os eventos mais distantes à CP, variando o tamanho do conjunto de dados reduzido (N_t);

Abordagem 3: Foi feita uma mistura entre as abordagens 1 e 2, em que utilizaram-se os dados mais próximos com uma parcela do conjunto reduzido (pp) de dados mais distantes à CP. Nesta abordagem, tanto o tamanho do conjunto reduzido (N_t) quanto à parcela de dados mais distantes (pp) foi variada.

Após a aplicação do procedimento de seleção de dados de acordo com a abordagem escolhida, foi realizado o teste dos dados com as Redes Neurais utilizando os conjuntos de dados reduzidos para o treinamento. Para isto, foi utilizada uma Rede Neural MLP como classificador com uma camada escondida com 10 neurônios. Como função de ativação, foi usada para todas as camadas a tangente hiperbólica. O resultado do treinamento da rede foi avaliado com validação cruzada do tipo *k-fold* com 10 *folds*. As medidas de avaliação utilizadas neste trabalho, foram baseadas nas medidas utilizadas por Freund (2018) sendo a probabilidade de detecção (P_D), a probabilidade de falso alarme (P_F) e o índice soma-produto (SP), este último sendo o critério utilizado para a escolha da Rede Neural durante o processo de treinamento a ser utilizada para o teste. Os hiperparâmetros da Rede Neural foram determinados de forma a maximizar o índice soma-produto. Este índice é calculado na Equação (1).

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

A partir destes dados, foram realizados os testes comparativos entre a rede neural treinada com todo o conjunto de dados e a rede neural treinada com os conjuntos de dados reduzidos por meio das medidas de desempenho citadas acima.

4. RESULTADOS E DISCUSSÃO

Após a geração das CP foram escolhidas, experimentalmente, Curvas Principais com 15 segmentos tanto para o padrão de sinal quanto de *background*. Tal escolha deve-se ao fato de CP com mais segmentos implicaram em maior tempo de desenvolvimento sem ganho de resultado de seleção significativo. Foram gerados gráficos de dispersão para cada padrão, apresentados nas Figuras 2 e 3 referentes, respectivamente, ao sinal de elétrons e ao ruído de fundo.

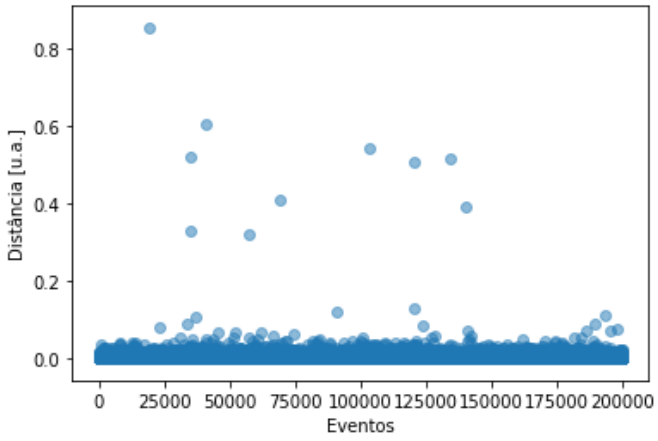

Figura 2. Gráfico de dispersão de distâncias dos eventos de sinal à sua respectiva CP.

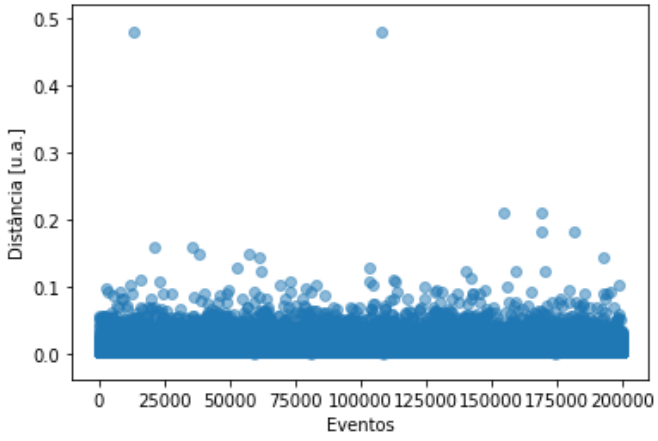

Figura 3. Gráfico de dispersão de distâncias dos eventos de ruído à sua respectiva CP.

Dada à grande concentração dos eventos nas baixas distâncias, foram feitos dois histogramas para cada padrão, tendo um corte para baixas e altas distâncias no ponto 0.03u.a. (unidade aritmética), de maneira a melhorar a distribuição dos eventos. Para efeitos de contextualização, o percentual de eventos nas distâncias acima do ponto de corte para o sinal foi de 0,0655% e para o ruído, o percentual foi de 0,8625%. O que mostra quase a totalidade dos eventos abaixo deste ponto de corte e que, um histograma contendo estes valores prejudicaria uma boa visualização da distribuição das distâncias. Os histogramas das baixas e altas distâncias para o padrão de sinal de elétrons são mostrados nas Figuras 4 e 5, respectivamente, enquanto os histogramas para as baixas e altas distâncias para o padrão de ruído de fundo estão contidos nas Figuras 6 e 7, respectivamente.

Os resultados de desempenho das Redes Neurais projetadas por meio das abordagens de seleção 1, 2 e 3 estão contidos, respectivamente nas Tabelas 2, 3 e 4, enquanto que na Tabela 5 são apresentados os resultados da Rede Neural para todo o conjunto de dados de treino, sem seleção de eventos via CP.

Observando os resultados gerados pelas Redes Neurais nas Tabelas 2, 3 e 4, e realizando uma análise comparativa com os resultados da Tabela 5, pode-se destacar que os valores atingidos pelo treinamento de conjunto de dados reduzidos utilizando o método de seleção atingiram números relativamente próximos dos resultados das Redes Neurais treinadas pelo conjunto de dados completo. Tal situação vale tanto para as taxas de detecção quanto para falso alarme, havendo, em alguns casos, resultados com valores ligeiramente superiores aos resultados do treinamento das Redes Neurais usando todo o conjunto de dados.

Por meio dos resultados quantitativos e enfatizando o comparativo dos resultados com os dados de teste, alguns pontos requerem destaque: (i) os resultados gerados pela Abordagem 1 apresentaram as menores taxas de falso alarme, contudo, apresentou as menores taxas de detecção dentre as 3 abordagens; (ii) os resultados das abordagens 2 e 3 apresentaram maiores valores de detecção, contudo, os valores de falso alarme foram maiores que os resultados obtidos pela abordagem 1 para alguns casos; (iii) mesmo, em alguns casos, a Abordagem 2 apresentando resultados de falso alarme melhores, não seria possível apontar de maneira significativa que seus resultados sejam superiores às demais abordagens.

Os resultados gerados pelas Figuras 2, 3, 4, 5, 6 e 7 mostram que os dados permanecem em maior concentração nas baixas distâncias. Também é observada uma grande discrepância entre os eventos com maior e menor distância. Tal situação sugere que tais eventos se apresentem como *outliers*, contudo, um estudo mais aprofundado necessita ser feito, uma vez que a seleção dos eventos mais distantes à CP contribuíram para a melhora dos resultados de detecção, como pode ser observado nos resultados das Abordagens 2 e 3 (Tabelas 3 e 4, respectivamente). Tal discrepância entre as distâncias também pode influenciar nos resultados no tocante ao desvio-padrão, portanto, realizar um estudo baseado em análise de agrupamentos mostra-se uma abordagem interessante, observando agrupamentos pelos segmentos das CP projetadas, sugerindo uma nova seleção de dados e, se possível, reduzir tais discrepâncias no mapeamento dos eventos.

Outro ponto a ser destacado é a redução no tempo de processamento das Redes Neurais proporcionada pela seleção de dados para o treinamento. Tal efeito foi mostrado em diferentes proporções dependendo da abordagem de seleção utilizada. Para efeitos de comparação, adotando-se o tamanho do conjunto N_t de 120.000 eventos, a Abordagem 1 registrou a maior taxa de redução, com valores chegando à 60% do tempo utilizando todo o conjunto de treinamento. Enquanto as Abordagens 2 e 3 apresentaram, respectivamente, números em torno de 33% e 37%. Tal cenário reforça que cada abordagem possui suas vantagens e desvantagens em relação às demais e que a escolha de uma em particular necessita levar em consideração cada um dos pontos apresentados. Contudo, os resultados observados mostraram-se promissores para um avanço nos estudos das CP como método de seleção de dados para um procedimento *online* com relação à diminuição da carga computacional sem perdas significativas da capacidade de generalização do modelo de aprendizagem.

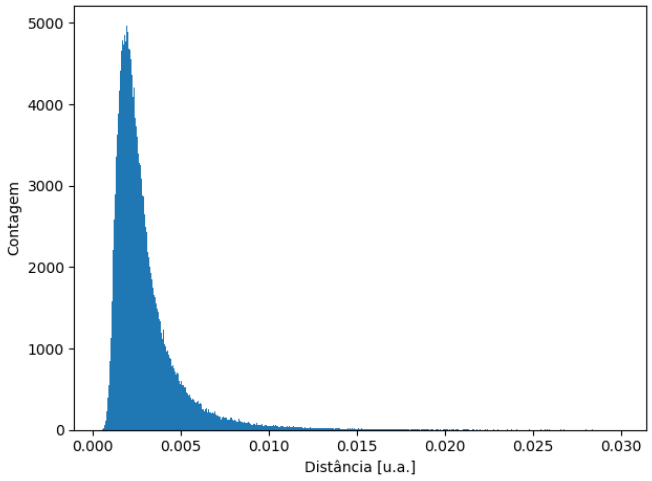

Figura 4. Histograma de distâncias para o sinal nas baixas distâncias (distâncias menores que 0.03u.a.).

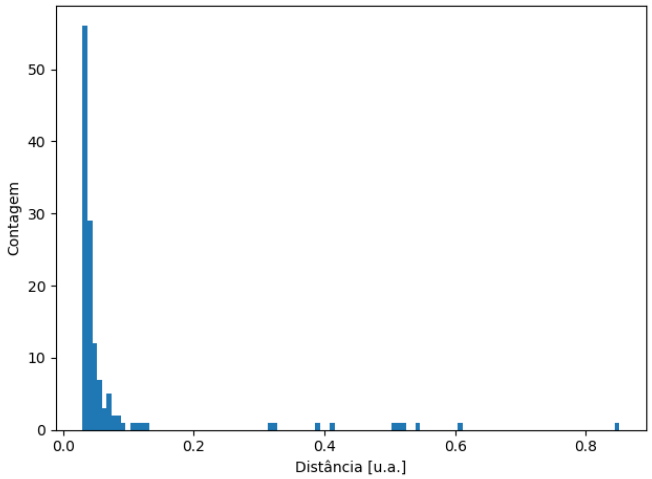

Figura 5. Histograma de distâncias para o sinal nas altas distâncias (distâncias maiores que 0.03u.a.)

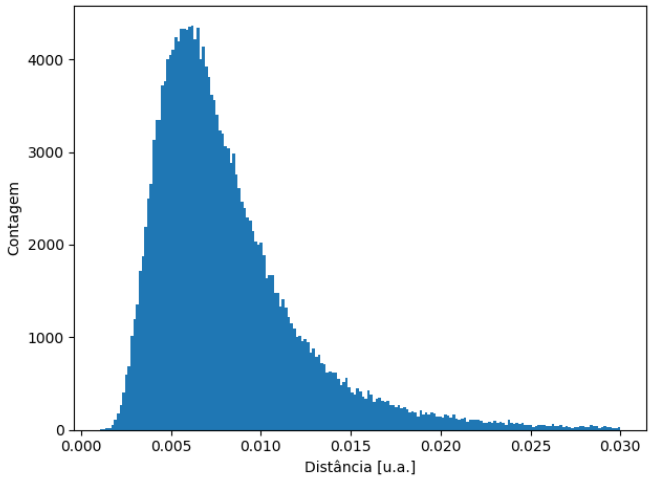

Figura 6. Histograma de distâncias para o ruído nas baixas distâncias (distâncias menores que 0.03u.a.)

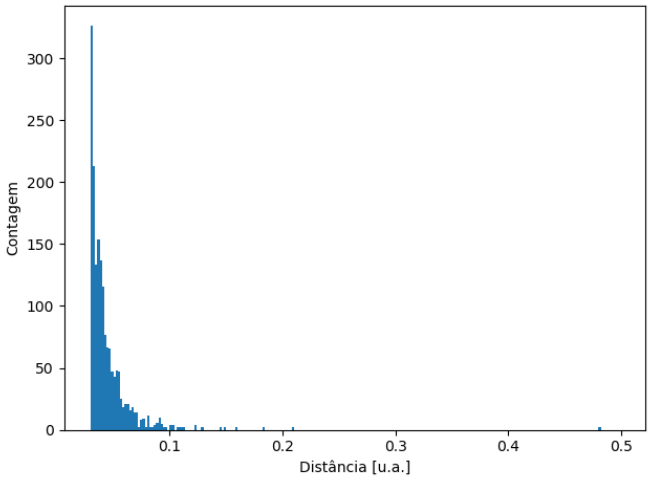

Figura 7. Histograma de distâncias para o ruído nas altas distâncias (distâncias maiores que 0.03u.a.)

Tabela 2. Resultados percentuais da Rede Neural (média ± desvio-padrão) utilizando a Abordagem 1 de seleção

		Treino		Teste	
N_t	SP_{max}	P_D	P_F	P_D	P_F
20.000	99,99 ± 0,02	99,99 ± 0,01	0,00 ± 0,04	94,19 ± 20,73	1,13 ± 9,57
40.000	99,99 ± 0,02	99,99 ± 0,02	0,00 ± 0,02	93,76 ± 22,58	1,19 ± 10,28
60.000	99,99 ± 0,19	99,99 ± 0,34	0,01 ± 0,17	93,36 ± 22,82	1,22 ± 10,21
80.000	99,99 ± 0,31	99,99 ± 0,52	0,00 ± 0,31	93,34 ± 23,85	1,28 ± 10,85
100.000	99,94 ± 1,23	99,95 ± 1,22	0,06 ± 1,75	94,05 ± 21,94	1,48 ± 11,52
120.000	99,84 ± 1,93	99,83 ± 2,27	0,11 ± 2,35	94,00 ± 21,24	1,75 ± 12,29
140.000	99,64 ± 3,41	99,68 ± 2,85	0,31 ± 4,50	95,17 ± 18,28	1,84 ± 12,42
160.000	99,30 ± 4,74	99,29 ± 4,63	0,51 ± 5,78	96,25 ± 15,44	2,11 ± 13,23
180.000	98,92 ± 6,21	99,05 ± 5,03	0,90 ± 8,01	97,21 ± 12,21	2,24 ± 13,39

Tabela 3. Resultados percentuais da Rede Neural (média ± desvio-padrão) utilizando a Abordagem 2 de seleção

		Treino		Teste	
N_t	SP_{max}	P_D	P_F	P_D	P_F
20.000	96,14 ± 12,33	96,81 ± 8,55	3,26 ± 15,78	98,36 ± 3,05	5,28 ± 19,09
40.000	96,37 ± 11,08	96,65 ± 8,96	2,92 ± 14,36	97,90 ± 4,12	3,35 ± 15,12
60.000	96,17 ± 11,28	96,53 ± 8,61	3,16 ± 15,10	98,08 ± 5,00	2,41 ± 12,36
80.000	96,37 ± 10,75	96,67 ± 8,85	2,99 ± 14,26	98,18 ± 5,77	2,13 ± 10,65
100.000	96,71 ± 10,69	97,18 ± 8,28	2,82 ± 14,06	98,22 ± 6,14	1,47 ± 8,60
120.000	97,08 ± 9,87	97,37 ± 8,15	2,43 ± 12,95	98,21 ± 6,60	1,22 ± 7,43
140.000	97,31 ± 9,47	97,67 ± 7,81	2,33 ± 12,46	98,19 ± 6,84	0,87 ± 5,92
160.000	97,51 ± 9,20	97,51 ± 9,20	2,16 ± 11,77	97,98 ± 7,87	0,92 ± 5,38
180.000	97,71 ± 8,95	97,98 ± 8,09	1,93 ± 11,26	98,21 ± 7,31	0,43 ± 3,49

Tabela 4. Resultados percentuais da Rede Neural (média ± desvio-padrão) utilizando a Abordagem 3 de seleção

			Treino		Teste	
N_t	pp	SP_{max}	P_D	P_F	P_D	P_F
80.000	0.9	98,99 ± 5,97	99,13 ± 6,47	0,86 ± 6,70	98,85 ± 6,62	2,78 ± 14,25
80.000	0.8	98,52 ± 7,27	98,76 ± 6,64	1,29 ± 8,79	99,00 ± 5,33	2,97 ± 14,54
80.000	0.7	98,08 ± 8,17	98,31 ± 7,59	1,61 ± 9,76	98,87 ± 5,15	2,83 ± 14,01
80.000	0.6	97,79 ± 8,78	97,89 ± 8,19	1,67 ± 10,35	98,79 ± 5,04	2,68 ± 13,59
120.000	0.9	98,79 ± 6,22	98,91 ± 6,69	1,04 ± 7,32	98,67 ± 6,84	3,05 ± 14,87
120.000	0.8	98,33 ± 7,69	98,60 ± 7,09	1,47 ± 9,55	98,68 ± 6,09	2,89 ± 14,32
120.000	0.7	98,11 ± 7,95	98,30 ± 6,97	1,57 ± 10,08	98,66 ± 6,02	2,63 ± 13,57
120.000	0.6	97,88 ± 8,48	98,16 ± 7,25	1,82 ± 10,79	97,96 ± 7,58	3,53 ± 15,22
160.000	0.9	98,45 ± 6,99	98,52 ± 7,36	1,24 ± 8,29	98,31 ± 7,58	3,10 ± 14,93
160.000	0.8	98,24 ± 7,83	98,36 ± 7,53	1,39 ± 9,21	98,46 ± 6,92	2,91 ± 14,44
160.000	0.7	98,11 ± 7,85	98,29 ± 7,21	1,57 ± 9,89	98,44 ± 6,70	2,64 ± 13,53
160.000	0.6	97,88 ± 8,67	98,29 ± 6,86	1,92 ± 11,37	98,24 ± 6,97	2,37 ± 12,60

Tabela 5. Resultados percentuais da Rede Neural (média ± desvio-padrão) utilizando todo o conjunto de desenvolvimento da CP

		Treino		Teste	
N_t	SP_{max}	P_D	P_F	P_D	P_F
200.000	98,01 ± 8,22	98,32 ± 6,82	1,76 ± 10,61	98,23 ± 7,37	1,70 ± 10,80

5. CONCLUSÕES

Para este trabalho, foi proposto um método de seleção inteligente de dados para treinamento de Redes Neurais com grande volume de dados em uma aplicação que envolve alta taxa de eventos utilizando Curvas Principais. Os resultados obtidos por meio de testes envolvendo o conjunto total de dados e os conjuntos de dados reduzidos por meio das abordagens de seleção propostas apresentaram um potencial do uso das CP na seleção de eventos para redução de processamento durante o ciclo de treinamento de máquinas.

Para avaliação dos resultados gerados pelo método, foi realizada uma simulação *offline* do procedimento realizado pelo *NeuralRinger*. Após análise dos resultados, foi observado que o desempenho de predição gerado pelos conjuntos de dados compactados apresentaram valores iguais ou melhores que os valores utilizando toda a massa de dados, viabilizando o avanço da técnica para teste em novos dados e implementação em *hardware*.

Para os próximos passos, tem-se por objetivos testar o método no algoritmo *NeuralRinger* e analisar os efeitos da seleção de dados, verificando, além dos resultados de detecção e falso alarme, se a seleção de eventos para o treinamento do modelo implicaria ou não, em tendências ao algoritmo. Tais estudos objetivam colaborar com melhorias computacionais ao método de filtragem *online* de elétrons de maneira que se possa manter o nível dos resultados e reduzir a carga computacional do sistema em um ambiente com elevados requisitos de processamento.

AGRADECIMENTOS

Agradecimentos às agências de fomento CAPES, CNPq, FAPEMIG, FAPERJ e à Universidade Federal de Lavras (UFLA) pelo aporte financeiro à esta pesquisa.

REFERÊNCIAS

Aad, G., Abajyan, T., Abbott, B., Abdallah, J., Khalek, S.A., Abdelalim, A.A., Abdinov, O., Aben, R., Abi, B., Abolins, M., et al. (2012). Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Physics Letters B*, 716(1), 1–29.

Al Nuaimi, E., Al Neyadi, H., Mohamed, N., and Al-Jaroodi, J. (2015). Applications of big data to smart cities. *Journal of Internet Services and Applications*, 6(1), 25.

Banfield, J.D. and Raftery, A.E. (1992). Ice floe identification in satellite images using mathematical morphology and clustering about principal curves. *Journal of the American Statistical Association*, 87(417), 7–16.

Chatrchyan, S., Khachatryan, V., Sirunyan, A.M., Tumasyan, A., Adam, W., Aguilo, E., Bergauer, T., Dragicevic, M., Erö, J., Fabjan, C., et al. (2012). Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Physics Letters B*, 716(1), 30–61.

Delicado, P. (2001). Another look at principal curves and surfaces. *Journal of Multivariate Analysis*, 77(1), 84–116.

Fan, J., Han, F., and Liu, H. (2014). Challenges of big data analysis. *National science review*, 1(2), 293–314.

Fong, S., Wong, R., and Vasilakos, A.V. (2015). Accelerated pso swarm search feature selection for data stream mining big data. *IEEE transactions on services computing*, 9(1), 33–45.

Freund, W.S. (2018). *Identificação de elétrons baseada em um calorímetro de altas energias finamente segmentado*. Ph.D. thesis, Universidade Federal do Rio de Janeiro.

Hasanin, T., Khoshgoftaar, T.M., Leevy, J., and Seliya, N. (2019). Investigating random undersampling and feature selection on bioinformatics big data. In *2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService)*, 346–356. IEEE.

Hastie, T. and Stuetzle, W. (1989). Principal curves. *Journal of the American Statistical Association*, 84(406), 502–516.

Jiang, S., Ferreira, J., and Gonzalez, M.C. (2017). Activity-based human mobility patterns inferred from mobile phone data: A case study of singapore. *IEEE Transactions on Big Data*, 3(2), 208–219.

Jin, X., Wah, B.W., Cheng, X., and Wang, Y. (2015). Significance and challenges of big data research. *Big Data Research*, 2(2), 59–64.

Kumar, S. and Singh, M. (2018). Big data analytics for healthcare industry: impact, applications, and tools. *Big Data Mining and Analytics*, 2(1), 48–57.

Onal, A.C., Sezer, O.B., Ozbayoglu, M., and Dogdu, E. (2017). 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)*, 2037–2046. IEEE.

Rong, M., Gong, D., and Gao, X. (2019). Feature selection and its use in big data: challenges, methods, and trends. *IEEE Access*, 7, 19709–19725.

Tibshirani, R. (1992). Principal curves revisited. *Statistics and computing*, 2(4), 183–190.

Verbeek, J.J., Vlassis, N., and Kröse, B. (2002). A k-segments algorithm for finding principal curves. *Pattern Recognition Letters*, 23(8), 1009–1017.