Adaptive online extreme learning machine by regulating forgetting factor by concept drift map

Hualong Yu, Geoffrey I. Webb

Elsevier's Neurocomputing 343 (2019)

Seminário: Ramon Durães

Apresentação

1) Introdução
2) Revisão Bibliográfica
    a) ELM
    b) Online Sequential ELM
    c) Online Sequential ELM with forgetting mechanism
    d) Outras abordagens
    e) Concept drift
3) Methods
    a) Concept Drift Map
    b) Estratégia para regulagem dinâmica do fator de esquecimento
    c) Descrição do algoritmo
4) Resultados
    a) Datasets
    b) Resultados diversos
    c) Discussões sobre os parâmetros dos modelo
    d) Tempo de execução
5) Conclusão
6) Referências Úteis

Introdução

Motivação: colocar modelos treinados em produção mostra a necessidade de modelos que aprendam em ambientes dinâmicos e:

- Aprendam com novas observações de forma “one pass”;
- Adaptem dinamicamente às variações na distribuição de dados

Outros autores desenvolveram modelos com fator de esquecimento, mas ele não se altera independente das mudanças na distribuição dos dados.

Introdução

- Os autores então desenvolvem uma estratégia para adaptar o fator de esquecimento dinamicamente baseado no Concept Drift Map.

Contribuições:

- Uma métrica de distância quantitativa entre duas distribuições construídas em um espaço contínuo de atributos (* **baseada em uma suposição sobre a distribuição**)
- Essa distância (que reflete a magnitude do concept drift) é associada com o fator de esquecimento, o que guia a adaptação do modelo

Revisão Bibliográfica
Notação

- **N** : número de observações
- **m** : número de classes
- **x_i** : vetor de entrada (n x 1)
- **t_i** : vetor de saída (m x 1)
- **L** : número de neurônios na camada escondida
- **T** : matriz (n x m) com todas as saídas
- **H** : saída da camada intermediária
- **β** : matriz de pesos da camada de saída

Revisão Bibliográfica

Extreme Learning Machines (ELM)

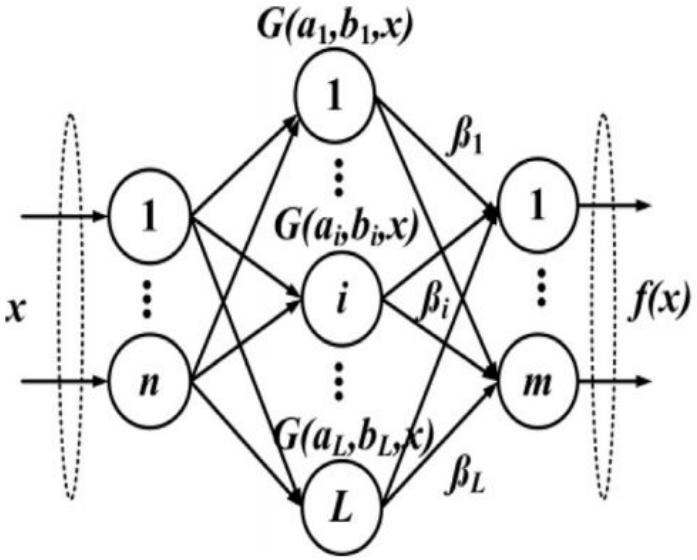


Fig. 1. The basic network structure of SLFN.

- MLP com pesos fixos da camada de entrada pra camada escondida: os demais são obtidos pelo sistema de equações:

$$\beta = \mathrm{H}^{\dagger}\mathrm{T} = \begin{cases} \mathrm{H}^{\mathrm{T}}(\mathrm{H}\mathrm{H}^{\mathrm{T}})^{-1}\mathrm{T}, \text{ when } N \le L \\ (\mathrm{H}\mathrm{H}^{\mathrm{T}})^{-1}\mathrm{H}^{\mathrm{T}}\mathrm{T}, \text{ when } N > L \end{cases} \tag{6}$$

Revisão Bibliográfica
Online Sequential ELM (OS-ELM)

Liang et al [34] - A fast and accurate online sequential learning algorithm for feedforward networks (2006)

- ELM que pode aprender incrementalmente em mini-batches;
- Obtém a mesma performance de uma ELM se treinada em todo o conjunto de dados;
- **P** depende apenas dos novos dados e **β** pode ser ajustado tanto para novos dados quanto para dados antigos (diminui tempo de cálculo dos parâmetros);

$$\beta^{(k+1)} = \beta^{(k)} + \mathrm{P}_{k+1}\mathrm{H}_{k+1}^{\mathrm{T}}(\mathrm{T}_{k+1} - \mathrm{H}_{k+1}\beta^{(k)}) \tag{7}$$

$$\mathrm{P}_{k+1} = \mathrm{P}_k - \mathrm{P}_k\mathrm{H}_{k+1}^{\mathrm{T}}(\mathrm{I} + \mathrm{H}_{k+1}\mathrm{P}_k\mathrm{H}_{k+1}^{\mathrm{T}})^{-1}\mathrm{H}_{k+1}\mathrm{P}_k \tag{8}$$

$$\mathrm{P}_0 = (\mathrm{H}_0^{\mathrm{T}}\mathrm{H}_0)^{-1} \tag{9}$$

Revisão Bibliográfica
OS-ELM with forgetting factor (SF-ELM)

Zhang and Wang [25] :Selective, Forgetting Extreme Learning Machine and its application to time series prediction (2011)

- OS-ELM lida eficientemente com o aprendizado online mas pode perder performance em ambientes não estacionários: “dá a mesma atenção a observações antigas e novas”;

$$\beta^{(k+1)} = (\mathrm{H}_k^{\mathrm{T}}\mathrm{H}_k + \mathrm{H}_{k+1}^{\mathrm{T}}\mathrm{H}_{k+1})^{-1}(\mathrm{H}_k^{\mathrm{T}}\mathrm{T}_k + \mathrm{H}_{k+1}^{\mathrm{T}}\mathrm{T}_{k+1}) \qquad (11)$$

- Adiciona um fator de esquecimento **η** nos termos relacionados às batches antigas

$$\beta^{(k+1)} = ((1-\eta)\mathrm{H}_k^{\mathrm{T}}\mathrm{H}_k + \mathrm{H}_{k+1}^{\mathrm{T}}\mathrm{H}_{k+1})^{-1}((1-\eta)\mathrm{H}_k^{\mathrm{T}}\mathrm{T}_k + \mathrm{H}_{k+1}^{\mathrm{T}}\mathrm{T}_{k+1}) \qquad (12)$$

- Quando:
  - **η = 0 :** volta a ser o OS-ELM
  - **η = 1 :** descarta todo o conhecimento antigo

Revisão Bibliográfica
Outras abordagens

[41] : Learning in the presence of concept drift and hidden contexts - Widmer and Kubat (1996)

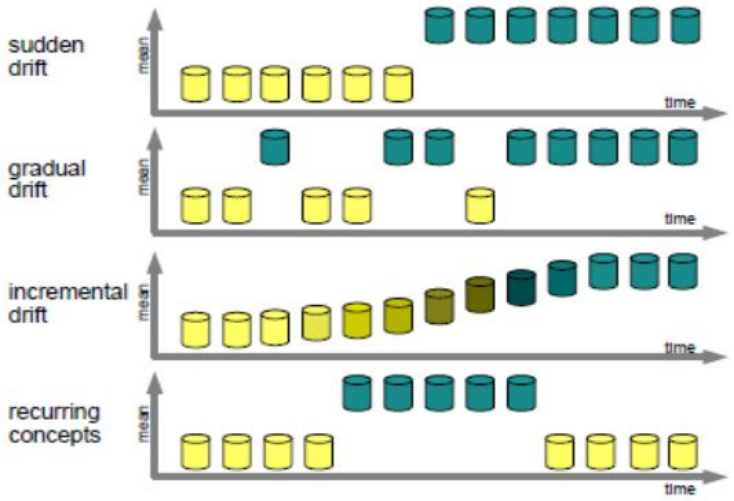
- Usa a magnitude do erro para definir o tamanho da sliding window
- Pontos negativos:
  - Só pode ser usado para streams de dados com atributos discretos
  - Só funciona com classificadores baseados em regras

[42] : Exponentially Weighted Moving Average (EWMA) charts for detecting concept drift - Ross et al. (2012)

- Chamado de EWMA-ELM no resto do trabalho;
- Monitora a média e desvio padrão da distribuição de entrada. Quando a variação passa de threshold, uma mudança é sinalizada.
- Desvantagens:
  - Requer que o modelo seja resetado (treinado do zero) caso seja detectada uma mudança
  - Funciona apenas para classificação binária

Revisão Bibliográfica
Concept Drift

- Variações na distribuição dos dados: distribuições “não-estacionárias”
- Vários tipos: abrupt drift, gradual drift, recurring drift ...



Métodos
Concept Drift Map

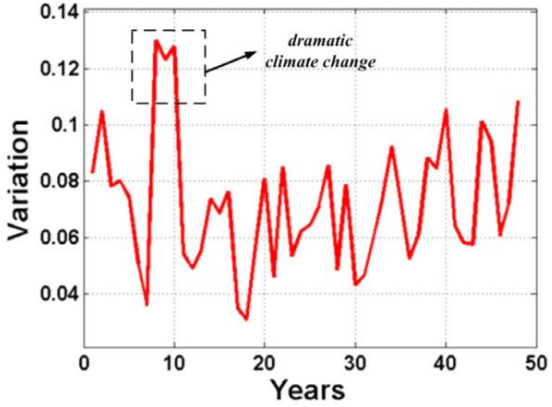


Fig. 2. An example of concept drift map for the NOAA data.

- Eixo vertical: magnitude da variação
- Eixo horizontal: bateladas de dados

Métodos

Concept Drift Map

$$
\begin{aligned}
s_{t,u} &= P(X_t > c) + P(X_u < c) \\
&= 1 - F_t(c) + F_u(c) \\
&= 1 - \frac{1}{2}\mathrm{erf}\left(\frac{c-\mu_t}{\sqrt{2}\sigma_t}\right) + \frac{1}{2}\mathrm{erf}\left(\frac{c-\mu_u}{\sqrt{2}\sigma_u}\right)
\end{aligned} \tag{19}
$$

$$
c = \frac{\mu_u\sigma_t^2 - \sigma_u\left(\mu_t\sigma_u + \sigma_t\sqrt{(\mu_t-\mu_u)^2 + 2(\sigma_t^2-\sigma_u^2)\log\left(\frac{\sigma_t}{\sigma_u}\right)}\right)}{\sigma_t^2 - \sigma_u^2} \tag{20}
$$

$$
\upsilon_{t,u} = 1 - s_{t,u} \tag{21}
$$

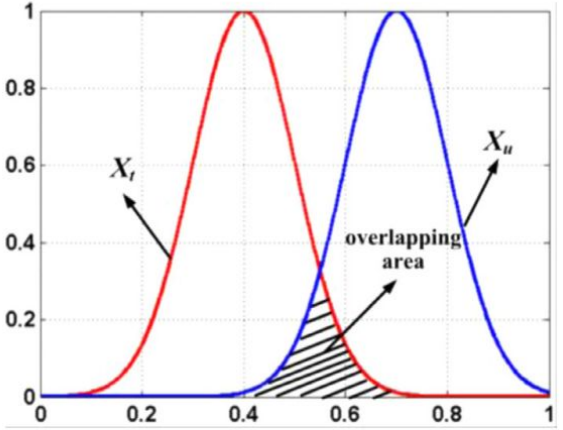


Fig. 3. Diagram illustrating the similarity $s_{t,u}$ calculation between two Gaussian distributions, where the area of overlap is considered their similarity.

Métodos

Concept Drift Map: Quantificando o concept drift

- Métrica conforma às condições de identidade, simetria, não negatividade e inequalidade triangular;
- Obtém-se um número por classe, por atributo, e então:
  - Calcula-se a média por atributo:

$$\upsilon_{t,u}(j) = \frac{1}{n}\sum_{l=1}^{n}(1 \dot{-} s_{t,u}(l, j)) \qquad (22)$$

  - Calcula-se a **média (AO-ELM$_{mean}$)** ou **máxima (AO-ELM$_{max}$)** por classe:

$$\upsilon_{t,u} = \underset{j\in\{1,2,\ldots,m\}}{\arg\max}\ \upsilon_{t,u}(j) \qquad (23)$$

$$\upsilon_{t,u} = \frac{1}{m}\sum_{j=1}^{m}\upsilon_{t,u}(j) \qquad (24)$$

Métodos

Regulagem Dinâmica do Fator de Esquecimento

- Define-se então um fator de esquecimento **η** proporcional à variação no drift **υ**:

$$\eta = 1 - \lambda - (1 - \lambda)\left(1/(\xi \exp(1))^{\upsilon}\right) \qquad (25)$$

- Em que:
  - **λ** : uma constante entre 0 e 1 que denota a “fixedly reserved percentage of old knowledge” (percentual fixo reservado de conhecimento anterior)
  - **ξ** : constante definida pelo usuário para regular a magnitude de conhecimento prévio que deve ser descartada. Definida empiricamente como 10 pelo resto do paper.
  -
- Se a batelada de dados é pequena, espera-se que a variação no drift **υ** seja mal estimada, logo **λ** deve ser grande.

Métodos

Pseudo código

globals

L: the number of hidden layer nodes

sig: the type of activation function

λ: the fixedly reserved percentage of old knowledge

ξ: the user-defined constant to regulate the old knowledge magnitude that should be forget

procedure Θ=AO-ELM(Ω=(Φ_0, Φ_1,..., Φ_z), *L*, *sig*, λ, ξ)

1. **collect** the first data chunk Φ_0;
2. **generate** the random weight and bias matrices *a* and *b* which are associated with *L* and *sig*;
3. **generate** H_0 by *a*, *b* and Φ_0;
4. **calculate** $P_0 = H_0^T H_0$;
5. **calculate** $\beta^{(0)}$ by Eq.(6);
6. **let** i=0;
7. **while** there are more data chunks, do the following:
8. $i=i+1$;
9. **receive** the i^{th} data chunk Φ_i;
10. **generate** H_i by *a*, *b* and Φ_i;
11. **acquire** predictive label subset $\Theta_i = H_i \beta^{(i-1)}$;
12. **calculate** the variation υ_i by Eq.(23) for the maximal variation or Eq.(24) for the average variation;
13. **calculate** η_i with υ_i, λ and ξ by Eq.(25);
14. **calculate** P_i with P_{i-1}, H_i and η_i by Eq.(18);
15. **update** $\beta^{(i)}$ with H_i, P_i and $\beta^{(i-1)}$ by Eq.(15);
16. **end while**

Fig. 4. Flow path description of AO-ELM algorithm.

Resultados
Datasets

- Foram utilizados 4 datasets públicos, dados reais e 4 datasets sintéticos
- Links no paper!

Table 1
Description about the used data sets. Except for electricity, which has one discrete attribute, all other attributes are continuous.

Data set	Number of attributes	Number of instances	Number of classes	Data chunk magnitude
Powersupply	2	29,928	24	72-168-240-720
HyperPlane	10	100,000	5	500-800-1000-1500
electricity	8	45,312	2	15-30-90-180
NOAA	8	18,159	2	15-30-90-365
SEA	3	10,000	2	30-50-80-100
spiral	2	80,000	2	200-400-600-800
Gaussian	2	80,000	2	400-800-1200-1600
checkerboard	2	20,000	2	50-100-150-200

Resultados
Parameters

Algumas escolhas de parâmetros foram feitas:

- SF-ELM : **η** empiricamente definido como 0.2
- EWMA-ELM : parâmetros recomendados pelo artigo original
- **AO-ELM** : **λ** variável (0.7, 0.5, 0.3, 0.1) de acordo com o tamanho da batelada de dados. **ξ** empiricamente mantido em 10;
- **L** : definido por grid search para todos os modelos.

Resultados

Exemplo de Tabela de Resultados

Table 5
0–1 loss on NOAA data set, where the value in each bracket denotes the p-value of t-test belonging to the corresponding algorithm compared with the best AO-ELM algorithm.

chunk size	OS-ELM	FOS-ELM	SF-ELM	EWMA-ELM	AO-ELM$_{mean}$	AO-ELM$_{max}$
15	$\mathbf{0.2067 \pm 0.0059}(2.37 \times 10^{-2})\blacktriangle$	$0.2583 \pm 0.0034(4.95 \times 10^{-4})\blacktriangledown$	$0.2187 \pm 0.0030(2.09 \times 10^{-1})$	$0.2219 \pm 0.0036(7.18 \times 10^{-2})$	0.2162 ± 0.0028	0.2188 ± 0.0029
30	$\mathbf{0.2049 \pm 0.0030}(1.28 \times 10^{-1})$	$0.2468 \pm 0.0053(8.79 \times 10^{-3})\blacktriangledown$	$0.2091 \pm 0.0051(1.33 \times 10^{-1})$	$0.2148 \pm 0.0032(5.14 \times 10^{-2})$	0.2082 ± 0.0044	0.2116 ± 0.0045
90	$0.2078 \pm 0.0040(5.30 \times 10^{-2})$	$0.2043 \pm 0.0040(1.74 \times 10^{-1})$	$0.2003 \pm 0.0029(3.62 \times 10^{-1})$	$0.2021 \pm 0.0043(2.55 \times 10^{-1})$	$\mathbf{0.1989 \pm 0.0043}$	0.2006 ± 0.0027
365	$0.2078 \pm 0.0036(3.76 \times 10^{-2})\blacktriangledown$	$0.1990 \pm 0.0051(8.46 \times 10^{-1})$	$0.1995 \pm 0.0047(7.44 \times 10^{-1})$	$0.1997 \pm 0.0037(5.63 \times 10^{-1})$	$\mathbf{0.1982 \pm 0.0036}$	$\mathbf{0.1987 \pm 0.0045}$

Resultados
Achados Diversos

- Os modelos com fator de esquecimento obtiveram melhor performance que o OS-ELM, reafirmando a eficácia desta abordagem em ambientes não estacionários.

- O tamanho da batelada de dados influencia bastante:
  - Aumenta / reduz o número de observações disponíveis para o treinamento do modelo;
  - Torna mais sutis / abruptas as mudanças detectadas pelos algoritmos;

Resultados
Tamanho da Batelada de Dados

- Exemplo do dataset NOAA para dados meteorológicos: a variação no drift só faz sentido para bateladas de tamanho 365, caso contrário, o drift excessivo é detectado comparando períodos que não são comparáveis.

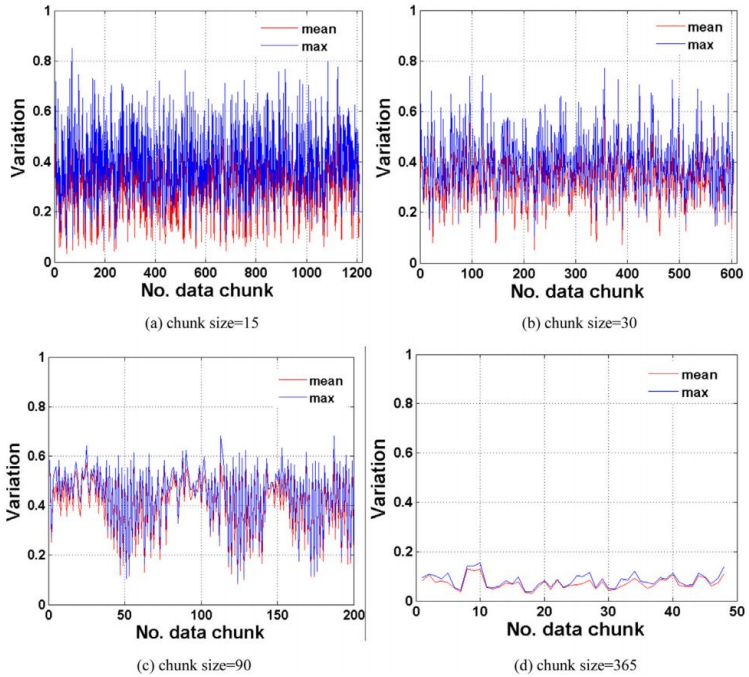


Fig. 5. Concept drift map on NOAA data set with different chunk sizes ((a): 15; (b): 30; (c): 90; (d): 365).

Resultados

Tamanho da Batelada e Magnitude de Variação **υ**

- Relação entre o tamanho da batelada e a magnitude da variação detectada.

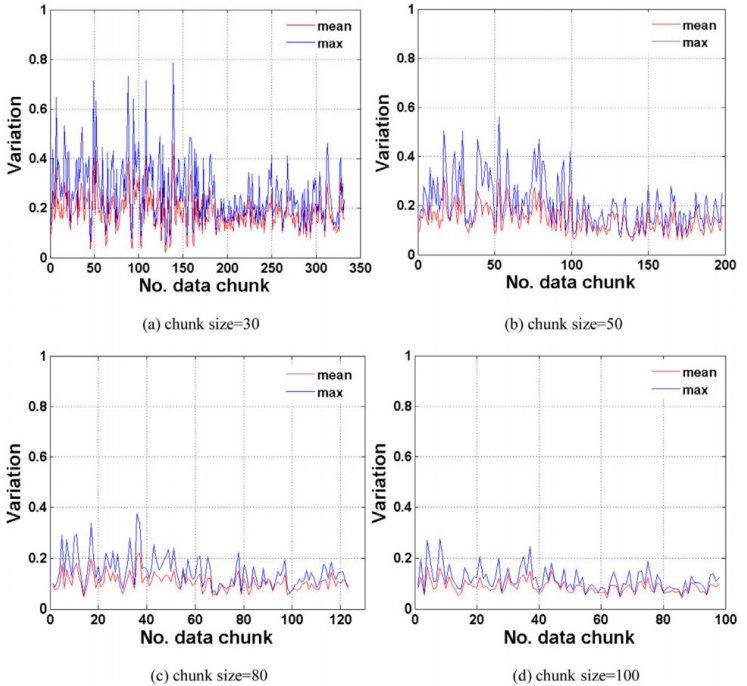


Fig. 6. Concept drift map on SEA data set with different chunk sizes ((a): 30; (b): 50; (c): 80; (d): 100).

Resultados
Achados Diversos

- A performance das duas variações do **AO-ELM** foi similar ou melhor à do SF-ELM para todas as bases de dados reais. Os autores argumentam que isso mostra a importância de se adaptar o fator de esquecimento à magnitude do drift (já que essa é a principal diferença entre eles).

- O algoritmo **AO-ELM** não se adapta bem ao concept drift recorrente. Isso porque as mudanças nos parâmetros são sempre monotônicas. Dessa forma o mecanismo de esquecimento não é ideal nestes casos.

Resultados

Tamanho da Batch e Parâmetro λ

- Sugestão dos autores para escolha do λ de acordo com o tamanho da batelada de dados:

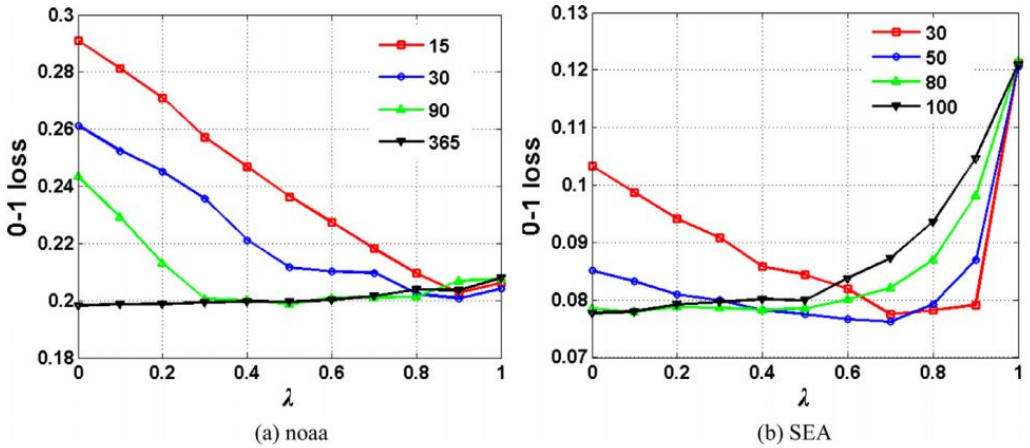


Fig. 7. Variation of 0–1 loss of AO-ELM$_{max}$ algorithm based on different λ values and different sizes of data chunk on NOAA data set (sub-graph (a)) and SEA data set (sub-graph (b)).

Resultados
Achados Diversos

- A performance de todos os modelos treinados foi bem pior nos conjuntos de dados com múltiplas classes.

- **AO-ELM$_{max}$** vs. **AO-ELM$_{mean}$** : Como esperado, o cálculo da média na variação das distribuições por classe atenua as mudanças. Os autores recomendam portanto a utilização do **AO-ELM$_{max}$**, mais sensível ao concept drift mas ainda robusto .

Resultados

Tempo de Execução

Table 10
Running time (s) on Powersupply data set.

chunk size	OS-ELM	FOS-ELM	SF-ELM	EWMA-ELM	AO-ELM$_{mean}$	AO-ELM$_{max}$
72	0.7972	0.9578	0.8784	2.1319	5.1543	5.0981
168	0.6802	0.7894	0.7176	1.2237	2.7643	2.7550
240	0.6880	0.7114	0.7238	1.5175	2.4149	2.4913
720	0.4789	0.4945	0.5249	1.0844	1.5116	1.5428

Table 11
Running time (s) on Hyperplane data set.

chunk size	OS-ELM	FOS-ELM	SF-ELM	EWMA-ELM	AO-ELM$_{mean}$	AO-ELM$_{max}$
500	2.6395	2.7534	2.7709	4.1997	6.5411	6.2681
800	2.0857	2.2152	2.1341	3.1966	4.8485	4.8485
1000	2.0218	2.0717	2.1296	3.7928	4.5427	4.5162
1500	1.9391	2.0093	2.0297	3.0946	4.0576	4.1028

Conclusão

- “Os experimentos mostram que o algoritmo proposto é uma solução eficaz e eficiente para a construção de modelos de aprendizado online em ambientes não estacionários.”
- Entretanto assume distribuição Gaussiana;
- Não lida bem com drift recorrente;
- Definição empírica do parâmetro **ξ**;
- Definição do parâmetro **λ** idealmente requer conhecimento do problema.

Trabalhos Futuros

- Explorar métodos que não suponham distribuição gaussiana;
- Buscar uma estratégia para definir os parâmetros ótimos automaticamente.

Referências Úteis

[25] Zhang and Wang - Selective, Forgetting Extreme Learning Machine and its application to time series prediction (2011) **SF-ELM**

[34] Liang et al - A fast and accurate online sequential learning algorithm for feedforward networks (2006) - **OS-ELM**

[41] : Learning in the presence of concept drift and hidden contexts - Widmer and Kubat (1996) - FLORA

[42] : Exponentially Weighted Moving Average (EWMA) charts for detecting concept drift - Ross et al. (2012) - **EWMA-ELM**