XLI CILAMCE
XLI Ibero-Latin American Congress on Computational Methods in Engineering
16-19 November, 2020 FOZ DO IGUAÇU PARANÁ - BRAZIL

Extensions and Improvements of the Extreme Learning Machine (ELM) Applied to Face Recognition

Iago G. L. Rosa[1], Ruan M. Carvalho[1], Alfeu D. Martinho[1,2], L. Goliatt[1,3], Raul F. Neto[1,4], Carlos C. H. Borges[1,4]

[1]*Graduate Program in Computational Modeling, Federal University of Juiz de Fora*
José Lourenço Kelmer street, s/n, University Campus, São Pedro, 36036-330, Juiz de Fora, Minas Gerais, Brazil
{iago.rosa, ruan.medina}@engenharia.ufjf.br, amartinho@ice.ufjf.br
{leonardo.goliatt, raulfonseca.neto}@ufjf.edu.br, cchborges@ice.ufjf.br
[2]*Department of Exact Sciences and Technology, Púnguè-Tete University*
EN106, Cambinde University Campus, Matundo, Tete, Tete province, Mozambique
[3]*Department of Computational and Applied Mechanics, Federal University of Juiz de Fora*
[4]*Department of Computer Science, Federal University of Juiz de Fora*

Abstract. Considering data science popularization in recent years, there has been an increase in using machine learning methods, such as artificial neural networks, in several research fields. The Extreme Learning Machine (ELM) method is a single-hidden layer feedforward network that stands out for its computational efficiency. It is reached mainly by adopting random weights initialization between the input layer and the hidden layer, as well as for the hidden neuron bias. This initialization allows the weights between the hidden and output layers to be determined analytically by calculating the pseudoinverse, avoiding the use of an iterative algorithm based on gradient descent. This work seeks to evaluate the performance of more sophisticated strategies for initializing the weights and fitting the machine's internal parameters. We proposed a performance analysis when replacing the pseudoinverse resolution step with the non-linear dual version of the linear least squares (LSDual), still little explored in the literature. We performed tests over face recognition databases in a classification approach. The results support the construction of an ELM model with a higher level of robustness regarding the quality of prediction. The use of more sophisticated random initialization coupled with LSDual solution reduced the testing error in all scenarios.

Keywords: *Face Recognition, Classification, Extreme Learning Machine, Dual Least Squares, Machine Learning*

1 Introdução

Com a popularização da ciência de dados nos últimos anos, houve um aumento no uso de inteligência artificial e métodos de aprendizado de máquina em vários campos de pesquisa. As redes neurais artificiais são uma parte essencial do aprendizado de máquina. Entre elas, estratégias *single-hidden layer feedforward networks* (SLFN), como descrito por Huang et al. [1], com topologia menos complexa, são uma das técnicas de aprendizado supervisionado mais usadas. Mesmo nestas topologias, há uma necessidade crescente de robustez e eficiência nos métodos propostos, além de uma melhor compreensão de seus padrões de aprendizagem.

O método Extreme Learning Machine (ELM), proposto por Huang et al. [2], é um SLFN que se destaca por sua eficiência computacional possibilitada, principalmente pela adoção de inicialização aleatória de pesos entre a camada de entrada e a camada oculta, assim como para os *bias* dos neurônios ocultos. Essa inicialização permite que os pesos entre a camada oculta e a camada de saída sejam determinados analiticamente, por técnicas de determinação de pseudoinversa, evitando o uso de um algoritmo iterativo baseado na *gradient descent*.

Desde que foi proposto em 2004, o método ELM mostrou-se robusto e vêm sendo alvo de diversas propostas de modificação. Tapson et al. [3], propõe uma expressão analítica para inicializar os pesos de entrada em um perceptron de várias camadas, que pode ser usado como o primeiro passo na síntese do ELM, já os pesos da camada oculta podem ser definidos com a estratégia analítica adequada. Gautam et al. [4] apresenta seis métodos OCC (*One-class classification*) e suas treze variantes baseadas em ELM e ELM sequencial *online* (OSELM). Tavares et al. [5] estuda a influência de métodos mais sofisticados de inicialização, em termos de desempenho e

complexidade. Por fim, Tavares et al. [6] apresenta uma técnica simples baseada em decomposição em valores singulares (SVD) que é capaz de indicar o número de neurônios da camada escondida que favoreça o baixo erro de treinamento e a baixa complexidade da máquina. Já no contexto de aplicação de classificação de imagens para reconhecimento facial, trabalhos recentes, como os de Jia et al. [7], Gao et al. [8] e Khalili Mobarakeh et al. [9], aplicam o ELM em testes sobre *benchmarks* com separações de treino e teste previamente estruturadas, aplicando pré-processamento aos dados precedendo as predições.

Nesse sentido, o presente trabalho avalia e compara o desempenho de diferentes funções de ativação e diferentes estratégias para a inicialização aleatória de pesos da rede em bases de dados de reconhecimento facial. Além disso, propõe-se uma análise de desempenho do método ao substituir a etapa de resolução de pseudoinversas pela solução da versão dual dos mínimos quadrados (LSDual), ainda pouco explorada na literatura.

2 Material e Métodos

2.1 Extreme Learning Machine (ELM)

O ELM proposto por Huang et al. [2] tem como objetivo realizar tarefas supervisionadas em conjuntos de dados. A ideia é que dado *N* distintas amostras $(\mathbf{x}_i, \mathbf{y}_i) \forall\, i \in \{1, 2, ..., N\}$ onde $\mathbf{x}_i = [x_{i1}, x_{i2}, \cdots, x_{in}] \in \mathbf{R}^n$ e $\mathbf{y}_i = [y_{i1}, y_{i2}, \cdots, y_{im}] \in \mathbf{R}^m$, um modelo de regressão ou classificação possa ser construído ajustado a esses dados. A estrutura do ELM, conforme sugerido por Xiao et al. [10], pode ser observado pela Figura 1.

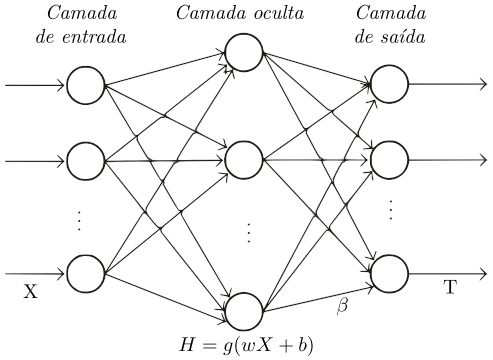

Figura 1. Representação do esquema de rede do ELM

De acordo com Huang et al. [2], o modelo considera um SLFN padrão e são escolhidos *L* neurônios na camada oculta de tal forma que é possível construí-la a partir de uma matriz de pesos aleatórios. O modelo matemático determinado pela eq. (1) considera as N amostas com erro nulo, ou seja, $\sum_{j=1}^{N} \|\mathbf{y}_j - \mathbf{t}_j\| = 0$. Isso significa que existe um $\boldsymbol{\beta}_i, \mathbf{w}_i$ e b_i de tal forma que:

$$\sum_{i=1}^{L} \boldsymbol{\beta}_i g\left(\mathbf{w}_i \cdot \mathbf{x}_j + b_i\right) = \mathbf{t}_j, \quad j = 1, \cdots, N \tag{1}$$

onde $\mathbf{w}_i = [w_{i1}, w_{i2}, \cdots, w_{in}]^T$ representa o *i*-ésimo neurônio na camada oculta e $i \in \{1, 2, ..., L\}$, $\boldsymbol{\beta}_i = [\beta_{i1}, \beta_{i2}, \cdots, \beta_{im}]^T$ é o peso da conexão da *i*-ésima camada oculta e o neurônio da camada de saída, e o b_i é o *bias* do *i*-ésimo do neurônio da camada oculta. Além disso, $g(\cdot)$ denota uma função de ativação. No presente trabalho, aplicam-se as funções de ativação dispostas na Tabela. 1. Além disso, a equação inicial descrita pela eq. (1) pode ser representada em notação matricial como $\mathbf{H}\boldsymbol{\beta} = \mathbf{T}$, onde:

$$\mathbf{H}\left(\mathbf{w}_1, \cdots, \mathbf{w}_L, b_1, \cdots, b_L, \mathbf{x}_1, \cdots, \mathbf{x}_N\right) = \begin{bmatrix} g\left(\mathbf{w}_1 \cdot \mathbf{x}_1 + b_1\right) & \cdots & g\left(\mathbf{w}_L \cdot \mathbf{x}_1 + b_L\right) \\ \vdots & \ldots & \vdots \\ g\left(\mathbf{w}_1 \cdot \mathbf{x}_N + b_1\right) & \cdots & g\left(\mathbf{w}_L \cdot \mathbf{x}_N + b_L\right) \end{bmatrix}_{N \times L} \tag{2}$$

$$\boldsymbol{\beta} = [\boldsymbol{\beta}_1 \cdots \boldsymbol{\beta}_L]_{L \times m}^T \quad \text{e} \quad \mathbf{T} = [\mathbf{t}_1 \cdots \mathbf{t}_N]_{N \times m}^T \tag{3}$$

Tabela 1. Funções de ativação usadas no ELM

	Nome	Função de ativação g
1	ReLU	$g(\mathbf{w}_i, b, \mathbf{x}_j) = \max_i (0, (\mathbf{w}_i \cdot \mathbf{x}_j + b))$
2	Sigmoide	$g(\mathbf{w}_i, b, \mathbf{x}_j) = \frac{1}{1+exp(-\mathbf{w}_i \cdot \mathbf{x}_j + b)}$
3	Tangente Hiperbólica (tanh)	$g(\mathbf{w}_i, b, \mathbf{x}_j) = \frac{1-exp(\mathbf{w}_i \cdot \mathbf{x}_j + b)}{1+exp(\mathbf{w}_i \cdot \mathbf{x}_j + b)}$

Conforme pode ser observado na Figura 1, $\mathbf{H}$ é referente a saída da camada oculta da rede neural e cada coluna de $\mathbf{H}$ diz respeito ao *i*-ésimo neurônio da saída da camada oculta em relação a entrada $\mathbf{x}_i$. Partindo da notação matricial do problema, o valor de $\boldsymbol{\beta}$ pode ser encontrado utilizando o método dos mínimos quadrados, ou seja, o valor ótimo do problema de minimização descrito pela eq. (4)

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^L} (\|\mathbf{H}\boldsymbol{\beta} - \mathbf{T}\|). \tag{4}$$

Dessa forma, o valor de $\boldsymbol{\beta}$ pode ser escrito como em eq. (5), onde $\mathbf{H}^\dagger$ é a pseudoinversa de $\mathbf{H}$.

$$\boldsymbol{\beta} = (\mathbf{H}^\mathbf{T}\mathbf{H})^{-1}\mathbf{H}^\mathbf{T}\mathbf{T} = \mathbf{H}^\dagger\mathbf{T} \tag{5}$$

2.2 Estratégias semi-randômicas para inicializações dos pesos das camadas

O método ELM considera uma inicialização aleatória para os pesos $\mathbf{w}_i \in \mathbb{R}^n$ e para os valores de viés $\mathbf{b}_i \in \mathbb{R}$. Ambos valores, conforme Huang et al. [2], são inicializados segundo uma distribuição uniforme (U), com função densidade de probabilidade descrita na eq. (6). Conforme sugerido por Tavares et al. [5], Tapson et al. [11], inicializações personalizadas podem gerar melhoras nos resultados. Visto isso, no presente trabalho, são consideradas 4 inicializações dos valores de $\mathbf{w}$ e $\mathbf{b}$, descritas a seguir:

Distribuição Uniforme: Se x é uma variável aleatória com distribuição uniforme *U(a, b)*, todo valor em $[a, b]$ tem a mesma probabilidade de ser observado. A função densidade de probabilidade de U é dada por:

$$f(x) = \begin{cases} \frac{1}{b-a}, \text{ se } a \leq x \leq b \\ 0, \text{ caso contrário.} \end{cases} \tag{6}$$

Distribuição Normal: Distribuição de probabilidades paramétrica simétrica em relação a sua esperança matemática (μ) com desvio-padrão (σ) e de função densidade de probabilidades expressa por:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}. \tag{7}$$

Nguyen Method: Método proposto por Nguyen and Widrow [12] considerando dois passos: i) $\mathbf{w} \approx U(-1, 1)$, ii) $\mathbf{w} = \frac{\beta \mathbf{w}}{\gamma}$; onde U é uma distribuição uniforme, $\beta = 0.7^{\frac{1}{L}}$, L é a quantidade de neurônios na camada de entrada e $\gamma = \sqrt{\sum_{i=1}^{L} w_i^2}$.

SCAWI: Método de nome *Controlled Activation Weight Initialization*, foi proposto por Drago and Ridella [13] e tem a inicialização dada por:

$$\mathbf{w} = \frac{1.3}{\sqrt{1 + Lv^2}} r_{i,j} \quad ; \tag{8}$$

onde $v = \sqrt{\sum_{i=1}^{L} w_i^2}$, L é a quantidade de neurônios na camada de entrada e $r \approx U(-1, 1)$.

2.3 Resolução do sistema: Pseusoinversa ou Mínimos Quadrados Dual

Conforme observado na eq. (5), a definição dos pesos da rede $\boldsymbol{\beta}$ refere-se à solução do método dos mínimos quadrados (LS) ou a um cálculo de pseudoinversa. Conforme Bauckhage [14], em bases que apresentam maior número de colunas do que de instâncias, soluções mais robustas podem ser obtidas com a aplicação da forma dual

dos mínimos quadrados (LSDual). Portanto, em bases de dados de imagens contendo grande número de pixels e número reduzido de imagens, o uso do LSDual pode ser promissor. A solução de $\boldsymbol{\beta}$ por meio da forma dual dos mínimos quadrados, considerando um termo de regularização $\lambda \in \mathbb{R}$, é descrita pela eq. (9):

$$\boldsymbol{\beta} = \boldsymbol{H}^T \left(\boldsymbol{K} + \lambda \boldsymbol{I}_n\right)^{-1} \boldsymbol{T} \quad ; \tag{9}$$

onde $\boldsymbol{K}$ é uma matriz *kernel*. Considera-se, para o mapeamento não linear dos dados, um *kernel* polinomial de grau 2 da forma $\boldsymbol{K} = (\boldsymbol{H}\boldsymbol{H}^T)^2$ com $\lambda = 0.2$.

2.4 Descrição das bases para reconhecimento facial

Foram selecionadas 3 bases de dados para a realização dos experimentos, dispostas na Tabela 2:

Yale Database: A base de dados **Yale**, de He et al. [15], Cai et al. [16], contém 165 imagens de faces de 15 pessoas. Cada pessoa contém 11 padrões diferentes como expressões, acessórios e luminosidade.

ORL Database: A base de dados **ORL**, de He et al. [15], Cai et al. [16], contém 400 imagens de faces de 40 pessoas. Cada pessoa contém 10 padrões diferentes como expressões, acessórios e luminosidade.

Extended Yale Face Database B: A base de dados **Yale B**, de Cai et al. [17], contém 2414 imagens de faces de 38 pessoas. Cada pessoa contém por volta de 64 padrões diferentes como expressões, acessórios e luminosidade.

As três bases de dados são utilizadas para testes de identificação de pessoas. O presente trabalho considera a utilização de 7 padrões diferentes para cada pessoa para compor o conjunto de treino nas duas primeiras bases de dados (referente a $\sim$64% dos dados na **Yale** e 70% na **ORL**). Já para a terceira base de dados, considera-se a utilização de 40 estados diferentes para compor o conjunto de treino. Os *benchmarks* já são disponibilizados[1] oferecendo 50 particionamentos treino-teste diferentes na configuração referida. Todos os particionamentos são utilizados para a avaliação de erros médios e de desvio padrão. Todas as imagens são apresentadas em tons de cinza, o que possibilita uma normalização de todos os pixel das bases de dados por 255.

Tabela 2. Resumo das bases, onde *h* é a altura e *c* o comprimento das imagens de entrada, e #treino, #teste, #características, #classes são os tamanhos dos conjuntos de treino, teste, características e classes.

Base de dados	h×c	#treino	#teste	#características	#classes
Yale	32x32	105	60	1024	15
ORL	32x32	266	134	1024	40
Yale B	32x32	1520	894	1024	38

2.5 Descrição dos experimentos computacionais

Os experimentos computacionais realizados visam investigar as alterações nas capacidades de identificação de pessoas à medida que diferentes funções de ativação ($FA \in \{$Relu, Sigmoide, tanh$\}$), inicializações da primeira camada ($HL_{ini} \in \{$Uniforme, Normal, Nguyen, SCAWI$\}$) e método de regressão ($\{$ELM-PInv, ELM-LSDual$\}$) eram variados na estrutura tradicional do ELM. Considerou-se 1000 neurônios na camada interna da rede do ELM em todas as execuções, como em Jia et al. [7]. O acompanhamento das qualidades das predições ocorreram por meio da acurácia de classificação das faces dada pela eq. (10):

$$\text{Acurácia} = \frac{1}{N} \sum_{i=1}^{N} I(f(x_i) = y_i) \quad ; \tag{10}$$

onde $f(x_i)$ é a classe prevista pelo modelo de uma instancia i e y_i é a verdadeira classe desta instância. Considere que $I(verdadeiro) = 1$ e $I(falso) = 0$. Os *benchmarks* disponibilizados já apresentam 50 divisões do conjunto

[1] http://www.cad.zju.edu.cn/home/dengcai/Data/FaceData.html

de dados em treinamento e teste, desse modo, calculou-se as médias e desvios da acurácia para cada uma das 3 bases de dados consideradas, com em Jia et al. [7], Gao et al. [8] e Khalili Mobarakeh et al. [9].

Os experimentos foram realizados em uma máquina 64 bits com Intel(R) Core(TM) i7-3632QM CPU, 2.20 GHz e 8 GB RAM usando a linguagem de programação Python3. Os dados utilizados na pesquisa, implementação computacional e resultados podem ser acessados por meio do repositório do github.

3 Resultados e Discussões

A Tabela 3 apresenta a acurácia média para as bases **Yale** e **ORL**. Dado o tamanho reduzido das bases, os métodos conseguiram se ajustar sem erro aos dados de treino*. Para ambos conjuntos de teste, a utilização do método ELM-LSDual com o conjunto $\{FA, HL_{ini}\}$ = {Sigmoide, Nguyen} apresentaram as melhores acurácias e baixo desvio padrão. Comparando a performance entre as duas bases, percebe-se que as acurácias para **ORL** foram maiores que para **Yale**. Esse comportamento pode ser proveniente da complexidade das imagens, ou com um conjunto de dados maior, a rede do ELM foi capaz de aprender melhor os padrões dos dados.

Tabela 3. Acurácia média no teste para as bases **Yale** e **ORL** para as diferentes combinações de parâmetros entre os métodos de regressão, função de ativação (FA) e inicialização dos pesos da camada oculta (HL_{ini}).

		Yale		**ORL**	
FA	HL_{ini}	ELM-PInv	ELM-LSDual	ELM-PInv	ELM-LSDual
Relu	Uniforme	0.7963 ± 0.0378	0.7717 ± 0.0406	0.9393 ± 0.0187	0.9527 ± 0.0165
	Normal	0.7843 ± 0.0460	0.7553 ± 0.0445	0.9400 ± 0.0229	0.9500 ± 0.0201
	Nguyen	0.7603 ± 0.0433	0.7617 ± 0.0428	0.8958 ± 0.0218	0.9132 ± 0.0203
	SCAWI	0.7767 ± 0.0484	0.5427 ± 0.0489	0.8870 ± 0.0201	0.6753 ± 0.0308
Sigmoide	Uniforme	0.8020 ± 0.0382	0.8017 ± 0.0385	0.8988 ± 0.0248	0.9315 ± 0.0252
	Normal	0.7473 ± 0.0405	0.7487 ± 0.0447	0.8505 ± 0.0255	0.9172 ± 0.0199
	Nguyen	0.7893 ± 0.0496	$\mathbf{0.8080 \pm 0.0457}$	0.9403 ± 0.0202	$\mathbf{0.9587 \pm 0.0166}$
	SCAWI	0.7910 ± 0.0437	0.5347 ± 0.0490	0.9313 ± 0.0212	0.5862 ± 0.0276
tanh	Uniforme	0.7710 ± 0.0403	0.7633 ± 0.0406	0.8335 ± 0.0338	0.9237 ± 0.0253
	Normal	0.7283 ± 0.0431	0.7297 ± 0.0406	0.7913 ± 0.0366	0.9193 ± 0.0220
	Nguyen	0.7820 ± 0.0468	0.7813 ± 0.0467	0.9302 ± 0.0209	0.9367 ± 0.0197
	SCAWI	0.7817 ± 0.0399	0.5443 ± 0.0486	0.9212 ± 0.0195	0.6347 ± 0.0344

*Os erros em relação ao treinamento para as bases **Yale** e **ORL** foram nulos todas as combinações FA e HL_{ini}, exceto para os testes com método de regressão LSDual e HL_{ini} = SCAWI, os quais apresentaram acurácia = 0.74 ± 0.04

A Tabela 4 apresenta a acurácia média nas etapas de treino e teste para a base **Yale B**. Os erros na etapa de treinamento foram consideravelmente reduzidos, o que mostra a capacidade do método em se ajustar ao conjunto de dados. No treinamento, o melhor conjunto $\{FA, HL_{ini}\}$ para o método ELM-PInv foi {Tangh, Normal}, seguido pelos conjuntos {tanh, Uniforme} e {Sigmoide, Normal}. Já no treinamento com o método EML-LSDual, uma série de combinações $\{FA, HL_{ini}\}$ apresentaram ajuste com acurácia máxima. Contudo, a análise dos resultados para o conjunto de teste demostram que os melhores resultados no treinamento podem caracterizar *overfitting*.

Os valores indicados em negrito na Tabela 4 representam os melhores resultados médios obtidos na etapa de teste. O método ELM-PInv alcançou seu melhor resultado com o conjunto $\{FA, HL_{ini}\}$ = {Relu, SCAWI}. É importante ressaltar que esse conjunto apresentou um dos maiores erros na etapa de treinamento, e mesmo assim obteve uma boa métrica para a generalização dos padrões no conjunto de teste. Apesar disso, o melhor resultado para a base **Yale B** foi alcançado pelo método EML-LSDual com o conjunto $\{FA, HL_{ini}\}$ = {Sigmoide, Nguyen}. Ressalta-se que esse foi o mesmo conjunto de parâmetros que gerou os melhores resultados para as bases **Yale** e **ORL**. Por fim, comparando a performance do método entre as três bases de dados, pode-se constatar que os níveis de erros médios decaíram com o aumento da base.

Tabela 4. Acurácia média no treino e teste para **Yale B** para as diferentes combinações de função de ativação (FA) e inicialização dos pesos da camada oculta (HL_{ini}). Resultados gerados por aplicação do método ELM-PInv.

		ELM-PInv		**ELM-LSDual**	
FA	HL_{ini}	Treino	Teste	Treino	Teste
Relu	Uniforme	0.9988 ± 0.0008	0.9026 ± 0.0104	1.0000 ± 0.0000	0.6053 ± 0.1040
	Normal	0.9986 ± 0.0008	0.9034 ± 0.0099	1.0000 ± 0.0000	0.4607 ± 0.1208
	Nguyen	0.9956 ± 0.0011	0.9563 ± 0.0069	0.9954 ± 0.0011	0.9597 ± 0.0067
	SCAWI	0.9958 ± 0.0011	$\mathbf{0.9578 \pm 0.0072}$	0.6969 ± 0.0236	0.6004 ± 0.0321
Sigmoide	Uniforme	0.9995 ± 0.0005	0.8901 ± 0.0098	1.0000 ± 0.0000	0.8852 ± 0.0285
	Normal	0.9997 ± 0.0004	0.8562 ± 0.0109	1.0000 ± 0.0000	0.9088 ± 0.0206
	Nguyen	0.9971 ± 0.0011	0.9563 ± 0.0069	0.9959 ± 0.0011	$\mathbf{0.9753 \pm 0.0055}$
	SCAWI	0.9973 ± 0.0010	0.9570 ± 0.0064	0.1811 ± 0.0238	0.1282 ± 0.0202
tanh	Uniforme	0.9997 ± 0.0004	0.8354 ± 0.0131	1.0000 ± 0.0000	0.9315 ± 0.0099
	Normal	0.9998 ± 0.0003	0.8051 ± 0.0117	1.0000 ± 0.0000	0.9409 ± 0.0084
	Nguyen	0.9970 ± 0.0010	0.9565 ± 0.0059	0.9962 ± 0.0011	0.9726 ± 0.0062
	SCAWI	0.9972 ± 0.0010	0.9576 ± 0.0063	0.6491 ± 0.0204	0.5486 ± 0.0299

Os resultados obtidos para as bases trabalhadas mostraram-se competitivos em relação aos resultados encontrados na literatura que utilizam metodologias similares de treino e teste e/ou que aplicam métodos com base no ELM, como em Jia et al. [7], Gao et al. [8], Liu et al. [18], e Khalili Mobarakeh et al. [9]. Com relação ao estado da arte disponibilizado pelo *benchmark*[1], os resultados se apresentaram resultados coerentes com as estratégias que não acoplam métodos de otimização de hiper-parâmetros ao método de classificação e/ou não realizam pré-processamento especial na base de dados (redução de dimensão ou convolução). Ambas estratégias podem ser acopladas ao método proposto neste trabalho, por exemplo por meio de abordagens descritas por Goliatt and Farage [19] ou Saporetti et al. [20], o que pode ser promissor como trabalho futuro

O método proposto pode ser encarado como uma alternativa mais abrangente ao método ELM tradicional de Huang et al. [2]. Além de propor uma inicialização modificada dos pesos da camada oculta, acopla um método de regressão dual que permite a inserção de matrizes *kernel* que podem contribuir para reconhecimento de características faciais, as quais podem ser entendidas como não lineares. Dessa forma, como mostrado por Khalili Mobarakeh et al. [9], as matrizes *kernel* mapeiam características não lineares apresentadas pelas faces e as classificam com maior eficiência. Sendo assim, acoplar o LSDual no ELM em problemas de classificação de imagens se torna uma boa alternativa. Em específico, nas bases de dados aplicadas nesse trabalho, a estratégia dual foi capaz de melhorar todos os resultados de testes obtidos quando aplicado com a inicialização de camada oculta apropriada.

Contudo, segundo Bauckhage [14], é importante ressaltar que a aplicação do LSDual é eficaz e recomendada quando não tivermos $n \gg m$, onde n é o número de características (colunas) e m é o número de observações (linhas). Caso contrário, a solução dual do problema pode se tornar extremamente mais custosa por conta do cálculo de $\boldsymbol{K}$ em eq. (9). Neste trabalho, observa-se que as bases de dados utilizadas mantiveram a mesma quantidade de características (1024) enquanto a quantidade de instâncias cresciam. Contudo, obtêm-se uma melhora nos resultados de teste. Mesmo com o aumento das bases, as matrizes resultantes no processo de ajuste de parâmetros se mantiveram com dimensões computacionalmente tratáveis com o aparato computacional disponível. Tal sucesso não seria diretamente replicável para bases com tamanhos extensivamente maiores.

4 Conclusão

O presente trabalho apresentou uma avaliação sobre o desempenho do método ELM potencializado com diferentes funções de ativação, estratégias mais sofisticadas para a inicialização aleatória de pesos da rede, e substituição a etapa de resolução de pseudoinversa pela solução da versão dual dos mínimos quadrados (LSDual). Os resultados mostraram que utilização de função de ativação Sigmoide, inicialização aleatória pela estratégia

de Nguyen, e ajuste de parâmetros internos por meio do LSDual foram capazes de melhorar a performance do método nos testes de todos os cenários apresentados. O método proposto se mostrou promissor frente aos resultados apresentados no *benchmark* utilizado. Como trabalhos futuros, pretende-se incluir outras funções de ativação para captar as não linearidades características deste problema, aplicar uma abordagem híbrida do ELM com estratégia evolutiva para otimização de hiper-parâmetros do modelo (número interno de neurônios e parâmetro de regularização do LSDual), adaptação das estratégias a um modelo de ELM multicamadas, realizar preprocessamento dos dados das imagens, assim como investigar variações de *kernel* não linear do LSDual.

Agradecimentos. Este trabalho foi financiado pela Universidade Federal de Juiz de Fora (UFJF), pela Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG), pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001, e pelo Programa de Formação de Professores de Educação Superior de Países Africanos (ProAfri) do Grupo Coimbra de Universidades Brasileiras (GCUB).

Declaração de autoria. Os autores confirmam que são as únicas pessoas responsáveis pela autoria deste trabalho, e que todo o material aqui incluído como parte deste artigo é de propriedade (e autoria) dos autores ou tem a permissão dos proprietários a serem incluídos aqui.

Referências

[1] Huang, G.-B., Wang, D. H., & Lan, Y., 2011. Extreme learning machines: a survey. *International journal of machine learning and cybernetics*, vol. 2, n. 2, pp. 107–122.
[2] Huang, G.-B., Zhu, Q.-Y., & Siew, C.-K., 2004. Extreme learning machine: a new learning scheme of feedforward neural networks. In *IEEE international joint conference on neural networks*, volume 2, pp. 985–990.
[3] Tapson, J., de Chazal, P., & van Schaik, A., 2014. Explicit computation of input weights in extreme learning machines. *ArXiv*, vol. abs/1406.2889.
[4] Gautam, C., Tiwari, A., & Leng, Q., 2017. On the construction of extreme learning machine for online and offline one-class classification—an expanded toolbox. *Neurocomputing*, vol. 261, pp. 126 – 143. Advances in Extreme Learning Machines (ELM 2015).
[5] Tavares, L. D., Saldanha, R. R., & Vieira, D. A., 2014a. Extreme learning machine with initialized hidden weight. In *2014 12th IEEE International Conference on Industrial Informatics (INDIN)*, pp. 43–47. IEEE.
[6] Tavares, L., Saldanha, R., & Vieira, D., 2014b. Seleção de número de neurônios de elms baseada em decomposição de valores singulares truncado.
[7] Jia, B., Li, D., Pan, Z., & Hu, G., 2015. Two-dimensional extreme learning machine. *Mathematical Problems in Engineering*, vol. 2015.
[8] Gao, Z., Zhang, G., Nie, F., & Zhang, H., 2017. Local shrunk discriminant analysis (lsda). *arXiv:1705.01206*.
[9] Khalili Mobarakeh, A., Cabrera Carrillo, J. A., & Castillo Aguilar, J. J., 2019. Robust face recognition based on a new supervised kernel subspace learning method. *Sensors*, vol. 19, n. 7, pp. 1643.
[10] Xiao, D., Li, B., & Mao, Y., 2017. A multiple hidden layers extreme learning machine method and its application. *Mathematical Problems in Engineering*, vol. 2017, pp. 1–10.
[11] Tapson, J., De Chazal, P., & van Schaik, A., 2015. Explicit computation of input weights in extreme learning machines. In *Proceedings of ELM-2014 Volume 1*, pp. 41–49. Springer.
[12] Nguyen, D. & Widrow, B., 1990. Improving the learning speed of 2-layer neural networks by choosing initial values of the adaptive weights. In *IJCNN International Joint Conference on Neural Networks*, pp. 21–26. IEEE.
[13] Drago, G. P. & Ridella, S., 1992. Statistically controlled activation weight initialization (scawi). *IEEE Transactions on Neural Networks*, vol. 3, n. 4, pp. 627–631.
[14] Bauckhage, C., 2015. Numpy / scipy recipes for data science: Kernel least squares optimization (1).
[15] He, X., Yan, S., Hu, Y., Niyogi, P., & Zhang, H.-J., 2005. Face recognition using laplacianfaces. *IEEE Trans. Pattern Anal. Mach. Intelligence*, vol. 27, n. 3, pp. 328–340.
[16] Cai, D., He, X., Han, J., & Zhang, H.-J., 2006. Orthogonal laplacianfaces for face recognition. *IEEE Transactions on Image Processing*, vol. 15, n. 11, pp. 3608–3614.
[17] Cai, D., He, X., & Han, J., 2007. Spectral regression for efficient regularized subspace learning. In *2007 IEEE 11th international conference on computer vision*, pp. 1–8. IEEE.
[18] Liu, C.-L., Zhang, C., & Wang, L., 2012. *Pattern Recognition: Chinese Conference, CCPR 2012, Beijing, China, September 24-26, 2012. Proceedings*, volume 321. Springer.
[19] Goliatt, L. & Farage, M. R. C., 2018. An elm with feature selection for estimating mechanical properties of lightweight aggregate concretes. In *IEEE Congress on Evolutionary Computation*, pp. 1–7.
[20] Saporetti, C. M., da Fonseca, L. G., & *et al.*, 2019. A lithology identification approach based on machine learning with evolutionary parameter tuning. *IEEE Geoscience and Remote Sensing*, vol. 16, n. 12, pp. 1819–1823.