https://doi.org/10.31513/linguistica.2020.v16nEsp.a18444
Edição Especial Comemorativa

APLICAÇÃO DO MÉTODO DE RECONHECIMENTO DE PADRÕES EM UM EXPERIMENTO LINGUÍSTICO COM CLASSIFICAÇÃO SUPERVISIONADA

APPLICATION OF THE PATTERN RECOGNITION METHOD IN A SUPERVISED CLASSIFICATION LINGUISTIC EXPERIMENT

Ali Kamel Issmael Junior[1]

Aline Gesualdi Manhães[2]

José Vicente Calvano[3]

RESUMO

Potenciais Relacionados a Eventos (ERP) são sinais elétricos biológicos sincronizados com estímulos sensoriais, cognitivos ou motores e medidos por eletroencefalógrafos (EEG). A técnica ERP permite a análise não invasiva das funções cerebrais. Com base nos resultados obtidos em um experimento linguístico de Soto (2014), este trabalho buscou a obtenção de cenários de classificação supervisionados para as classes propostas no trabalho mencionado, bem como o estudo comparativo e a discussão dos resultados de classificação encontrados, utilizando a metodologia proposta por Webb (2012).

PALAVRAS-CHAVE: ERP, EEG, Reconhecimento de Padrões, Computação Linguística.

ABSTRACT

Event-Related Potentials (ERP) are biological electrical signals synchronized with sensory, cognitive or motor stimuli and measured by electroencephalographs (EEG). ERP technique allows non-invasive analysis of brain functions. Based on the results obtained in Soto (2014) linguistic experiment, this work sought to obtain supervised classification scenarios for the classes proposed in the mentioned work, as well as the comparative study and the discussion of the classification results found, using the methodology proposed by Webb (2012).

KEYWORDS: ERP, EEG, Pattern Recognition, Linguistic Computations.

[1] Centro Federal de Educação Tecnológica Celso Suckow da Fonseca (CEFET-RJ). Mestre em Engenharia Elétrica. Contato: alikamel@ig.com.br.

[2] Centro Federal de Educação Tecnológica Celso Suckow da Fonseca (CEFET-RJ). Professor do Programa de Pós-Graduação em Engenharia Elétrica. Contato: alinegesualdi@gmail.com.

[3] Universidade Federal do Rio de Janeiro (UFRJ-COPPE). Doutor em Engenharia Elétrica; Instituto de Pesquisas da Marinha. Contato: jvcalvano@gmail.com.

I. INTRODUÇÃO

Os Potenciais Relacionados a Eventos (ERP) são voltagens elétricas associadas a uma resposta neurofisiológica induzida por um evento ou estímulo externo. Os ERPs são obtidos por meio da Eletroencefalografia (EEG), que é um aparato não invasivo sensível o suficiente para medir pequenos potenciais elétricos no couro cabeludo humano, como resultado da estimulação por eventos sensoriais, cognitivos ou motores.

Este estudo utiliza os dados experimentais do ERP de Soto (2014) obtidos considerando funções cognitivas subjacentes do componente ERP mensurados em palavras-alvo em contextos sentenciais e de pares de palavras em língua portuguesa, para aplicações em neurolinguística.

A partir desses dados experimentais de ERP, e do uso de ferramentas específicas de computador EEGLAB® (EEGLAB®, 2016) e ERPLAB® (ERPLAB®, 2016), (LOPEZ-CALDERÓN e LUCK, 2014), com base no software MATLAB® (MATLAB®, 2016), é possível tratar esses dados experimentais para investigar se existem parâmetros específicos de reconhecimento para os sinais ERP relacionados a cada tipo de estímulo. O tratamento desses sinais ERP envolve o estudo de técnicas de processamento digital de sinais aplicadas com a teoria de reconhecimento de padrões.

Dessa forma, o objetivo deste trabalho é investigar, aplicando a metodologia de reconhecimento de padrões proposta por Webb (2012), nos resultados do experimento ERP de Soto (2014), se é possível obter bons cenários de classificação para os cenários definidos. No trabalho anterior (ISSMAEL Jr., GESUALDI e CALVANO, 2017), o estímulo para as épocas não rotuladas (classificação não supervisionada e métodos de agrupamento) alcançou precisões muito próximas da equiprobabilidade, indicando que o uso delas não é adequado para classificar os dados de Soto (2012). Neste artigo, as épocas rotuladas (classificação supervisionada) serão abordadas, sendo verificada a consistência das sentenças e classes de palavras propostas por Soto (2012), através da extração de atributos dos sinais EEG e ERP para cada época e sua associação com as classes previamente identificadas.

Este estudo é inovador na área de Neurolinguística, uma vez que, pelo menos até o momento, não existem trabalhos similares publicados anteriormente sobre o assunto encontrados em bancos de dados de pesquisa como: IEEExplorer; Web of Science; Elesevier e Spring. Os resultados abrem a possibilidade de analisar sinais de indivíduos com esta metodologia ERP associada ao Reconhecimento de Padrões, com a possível aplicação desse tipo de análise em ferramentas de diagnóstico, avaliação de aprendizagem de línguas, entre outros.

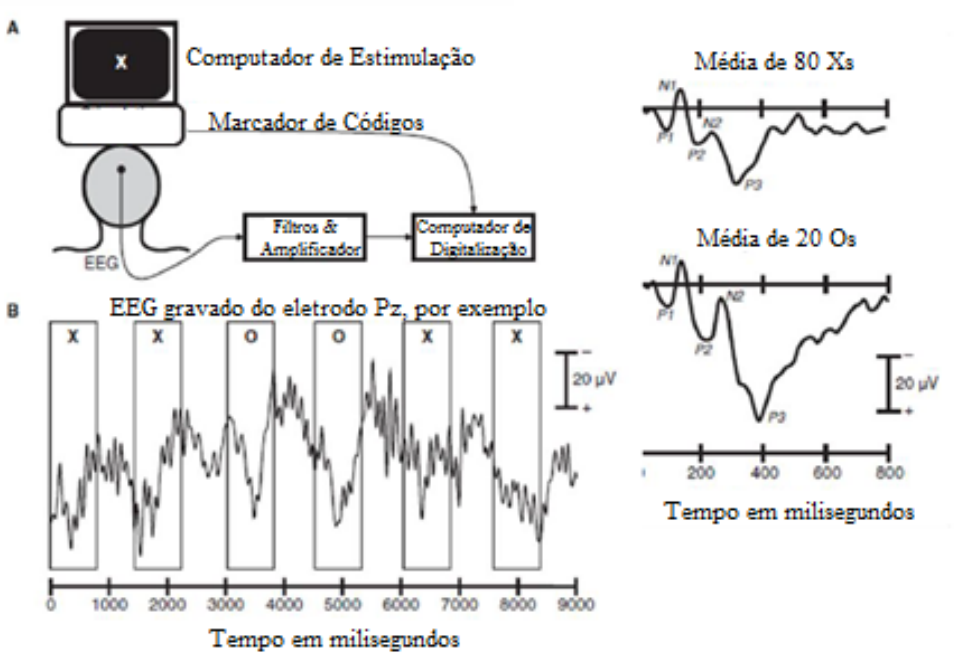
II. REFERÊNCIAS TEÓRICAS

A. Teoria sobre EEG e ERP

Os sinais ERPs são obtidos por procedimento de EEG que mede a atividade elétrica do cérebro ao longo do tempo, por meio de eletrodos colocados no couro cabeludo. O EEG reflete milhares de processos cerebrais simultaneamente em andamento em pontos específicos distribuídos em áreas específicas e Regiões de Interesse (ROI) do couro cabeludo, dependendo do alvo cognitivo da pesquisa. Esses ROI estão mais relacionados com o processo de análise cognitiva do que com a coleta de dados relacionada aos eletrodos individuais. A resposta do cérebro a um único estímulo ou evento de interesse não é geralmente visível no registro de EEG de um único teste. A técnica de ERP consiste em uma amplificação de sinal que soma e calcula a média de épocas especificamente bloqueadas pelo tempo, que são replicações de um estímulo, e idealmente pode apresentar uma relação sinal / ruído (SN) menor que as formas de onda originais. A relação sinal-ruído (SN) é uma medida de qualidade para sinalizar o processamento, sendo definida como a relação entre a potência do sinal e a potência do ruído (GESUALDI e FRANÇA, 2011).

O sinal EEG é registrado como um sinal contínuo, e a apresentação do estímulo é marcada em seu início, onde as tarefas de detectar e marcar o sinal apresentam alguma dificuldade. O sinal bruto é geralmente filtrado para baixas frequências (por exemplo, passagem alta de 0,01 Hz) e amplificado. Depois disso, um computador, ou uma caixa de disparo conectada separadamente, marca um código digital ou uma largura de pulso no sinal gravado, permitindo a marcação no sinal EEG contínuo do início exato do estímulo e o tipo de estímulo mostrado. Um exemplo ilustrativo de um experimento é apresentado esquematicamente na Figura 1: os sujeitos viram muitos “X”, alternados com “O”s. Os fragmentos, chamados épocas, relacionados ao evento são calculados para cada eletrodo de modo a amplificar a resposta e filtrar o ruído proveniente de outra atividade neurofisiológica ou a interferência de equipamentos elétricos. Estas respostas médias, os ERPs, podem agora ser comparados e caracterizados em termos de amplitude (em μV) - o pico da onda - e latência (em ms) - o tempo no qual a onda atinge seu pico [1]. A figura 1 mostra um esquema simplificado para o experimento de ERP.

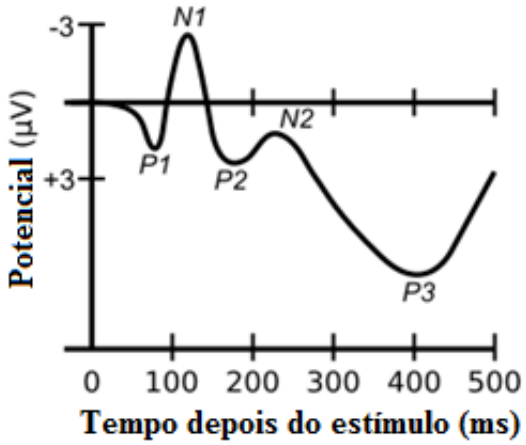
Figura 1 - Exemplo de experimento EEG / ERP com épocas "X" e "O" (LUCK, 2014)



Os sinais ERP resultantes do experimento apresentam uma série de deflexões de tensão positiva e negativa, que estão relacionadas a um conjunto de componentes subjacentes chamados componentes de ERP. Os componentes usuais do ERP são referidos por uma letra (N / P) indicando polaridade (negativa / positiva), seguida por um número indicando a latência em milissegundos ou a posição ordinal do componente na forma de onda (LUCK, 2014).

Por exemplo, um pico negativo que é o primeiro pico substancial na forma de onda e, geralmente, ocorre cerca de 100 milissegundos depois que um estímulo é apresentado, é frequentemente chamado de N100 (indicando que sua latência é de 100 ms após o estímulo e que é negativo) ou N1 (indicando que é o primeiro pico e é negativo); geralmente é seguido por um pico positivo, geralmente chamado de P200 ou P2. As latências declaradas para os componentes do ERP costumam ser bastante variáveis. Por exemplo, o componente P300 pode exibir um pico em qualquer lugar entre 250ms - 700ms (WOODMAN, 2010). Na Figura 2 abaixo, um exemplo de um sinal de forma de onda ERP com esses componentes pode ser visto.

Figura 2 - Um exemplo de gráfico de forma de onda ilustrativa fictícia mostrando vários componentes de ERP como P1 (P100), N1 (N100), P2 (P200), N2 (N200) e P3 (P300) (LUCK, 2014).



Conforme descrito por Woodman (2010), um componente ERP pode ser simplesmente definido como uma das ondas componentes da forma de onda mais complexa do ERP. Os componentes do ERP são definidos por sua polaridade (voltagem positiva ou negativa), tempo, distribuição do couro cabeludo e sensibilidade às manipulações de tarefas. Diferentes nomenclaturas de componentes de ERP enfatizam diferentes aspectos dessas características definidoras e fornecem um ponto de partida para revisões de literatura.

Com relação ao objetivo deste trabalho (propriedades específicas da linguagem), Soto (2014) indica que, de fato, as metodologias de ERP trouxeram muitas evidências para mostrar que informações linguísticas muito detalhadas têm um efeito imediato no processamento de fluxos. Soto (2014) também indica que o componente N400 do sinal ERP pode ser influenciado por variáveis linguísticas estritas. Sendo assim, para este estudo, três parâmetros do sinal ERP foram extraídos do experimento como atributos para os classificadores, por terem uma boa possibilidade de serem influenciados pelos estímulos propostos. Estes parâmetros foram:

a) Amplitude média entre duas latências fixas - Amplitude média de pico em um intervalo de tempo de ERP;
b) Amplitude de pico - a amplitude máxima de pico em um intervalo de tempo de ERP; e
c) Latência do pico - o valor do tempo onde ocorre a amplitude máxima do pico.

O uso destes parâmetros será melhor explanado adiante.

B. O Experimento ERP de Soto (2014)

A partir das variáveis propostas (SOTO, 2014), 4 condições de estímulo e uma condição de controle foram estabelecidas para o Experimento/Tarefa de "Sentenças": (i) contexto de apoio congruente (CSC): "Sem capacete, João dirige a moto como louco"; (ii) Contexto Congruente Não Suportivo (CNSC): "Todos os dias, João dirige um moto feito louco"; (iii) Contexto de Suporte Incongruente (ISC): "Sem capacete, João dirige a pêra como um louco"; (iv) Contexto Não Suportivo Incongruente (INSC): "Todos os dias, João dirige um pera feito louco"; e (v) Sentença de Controle. Para o Experimento/Tarefa de pares de "Palavras", Soto (2014) propôs 2 condições de estímulo e duas condições de controle: (i) Relação Semântica Associativa (ASR): "ÔNIBUS moto"; (ii) Relação Sintática e Semântica (SSR): "CAPACETE moto"; (iii) Controle 1 - Par Não Relacionado (UR): "FACA nuvem; e Controle 2 - Par com Alvo de Pseudopalavra (PW): "Garufa CARRO".

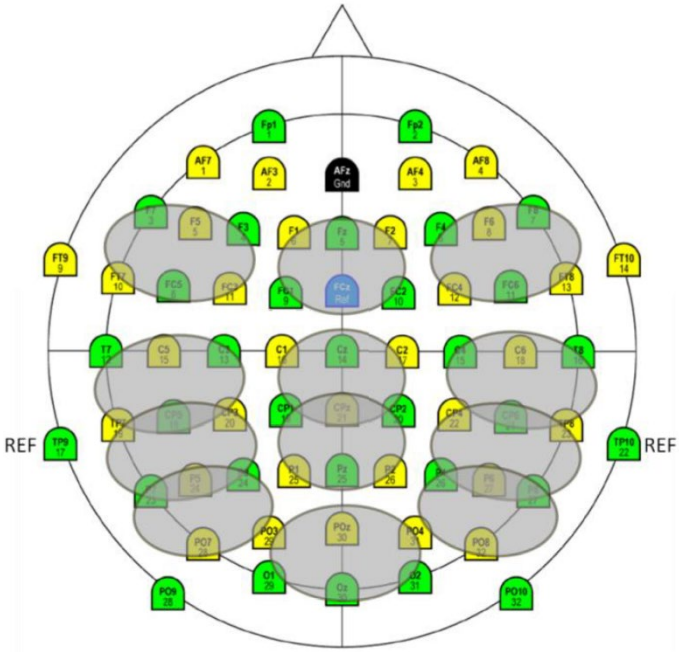
Quanto à montagem experimental, participaram do estudo 21 estudantes universitários (sendo 11 mulheres), distribuídos uniformemente em 4 versões, com idade média de 22 anos, todos destros, com visão normal ou corrigida para o normal. Os julgamentos dos participantes foram registrados com eles pressionando com um dos dois dedos da mão direita, um botão vermelho ou verde em uma caixa de botão. A posição dos botões verde e vermelho, destinados a respostas SIM e NÃO, foi trocada para cada participante. A Figura 3 ilustra o experimento:

Figura 3 – O Experimento de Soto (2014) (SOTO, 2014)

As Regiões de Interesse (ROIs) do couro cabeludo foram definidas da seguinte forma: ao longo da linha média foram: Frontal (F1-ch34, F2-ch60, FC1-ch7, FC2-ch27, FCz-ch38 e

Fz-ch2) ; Central (C1-ch39, C2-ch56, CP1-ch11 CP2-ch21, CPz-ch52 e Cz-ch22), Parietal (CP1-ch11, CP2-ch21, CPz-ch52, P1-ch43, P2-ch51 e Pz -ch12) e Occipital (O1-ch15, O2-ch17, Oz-ch16, PO3-ch46, PO4-ch48 e POz-ch47). No hemisfério esquerdo, eles eram Frontal (F3-ch3, F5-ch35, F7-ch4, FC3-37, FC5-ch6 e FT7-ch36); Central (C3-ch8, C5-ch40, CP3-ch42, CP5-ch10, T7-ch9 e TP7-ch41), Parietal (CP3-ch42, CP5-ch10, P3-ch13, P5-ch40, P7-14 e TP7 -ch41) e Occipital (P3-ch13, P5-ch44, P7-ch14, PO3-ch46 e PO7-ch45). E no hemisfério direito, foram: Frontal (F4-ch28, F6-ch59, F8-ch29, FC4-ch57, FC6-ch26 e FT8-ch58); Central (C4-ch23, C6-ch55, CP4-ch53, CP6-ch20, T8-ch24 e TP8-ch54), Parietal (CP4-ch53, CP6-ch20, P4-ch18, P6-ch50, P8-ch19 e TP8 -ch54) e Occipital (P4-ch18, P6-ch50, P8-ch19, PO4-ch48 e PO8-ch49). Essa disposição das ROI é apresentada na Figura 4.

Figura 4 - Definição das ROI baseada na proximidade anatômica (SOTO, 2014)



Para obter os sinais ERP para cada ROI, é necessário adicionar a contribuição de cada canal de eletrodo relacionado com a Região e obter a mídia aritmética. Assim, considerando a distribuição do eletrodo do experimento, o sinal ERP para cada ROI é obtido pelas seguintes equações:

$$\text{Linha Média Frontal (ch63)} = (\text{ch2} + \text{ch7} + \text{ch27} + \text{ch34} + \text{ch38} + \text{ch60}) / 6 \quad (1)$$

$$\text{Linha Média Central (ch64)} = (\text{ch11} + \text{ch21} + \text{ch22} + \text{ch39} + \text{ch52} + \text{ch56}) / 6 \quad (2)$$

$$\text{Linha Média Pariental (ch65)} = (\text{ch11} + \text{ch12} + \text{ch21} + \text{ch43} + \text{ch51} + \text{ch52}) / 6 \quad (3)$$

$$\text{Linha Média Occipital (ch66)} = (\text{ch15} + \text{ch16} + \text{ch17} + \text{ch46} + \text{ch47} + \text{ch48}) / 6 \quad (4)$$

Lado Esquerdo Frontal (ch67) = (ch3 + ch4 + ch6 + ch35 + ch36 + ch37) / 6 (5)

Lado Esquerdo Central (ch68) = (ch8 + ch9 + ch10 + ch40 + ch41 + ch42) / 6 (6)

Lado Esquerdo Pariental (ch69) = (ch10 + ch13 + ch14 + ch41 + ch42 + ch44) / 6 (7)

Lado Esquerdo Occipital (ch70) = (ch13 + ch14 + ch44 + ch45 + ch46) / 5 (8)

Lado Direito Frontal (ch71) = (ch26 + ch28 + ch29 + ch57 + ch58 + ch59) / 6 (9)

Lado Direito Central (ch72) = (ch20 + ch23 + ch24 + ch53 + ch54 + ch55) / 6 (10)

Lado Direito Pariental (ch73) = (ch18 + ch19 + ch20 + ch50 + ch53 + ch54) / 6 (11)

Lado Direito Occipital (ch74) = (ch18 + ch19 + ch48 + ch49 + ch50) / 5 (12)

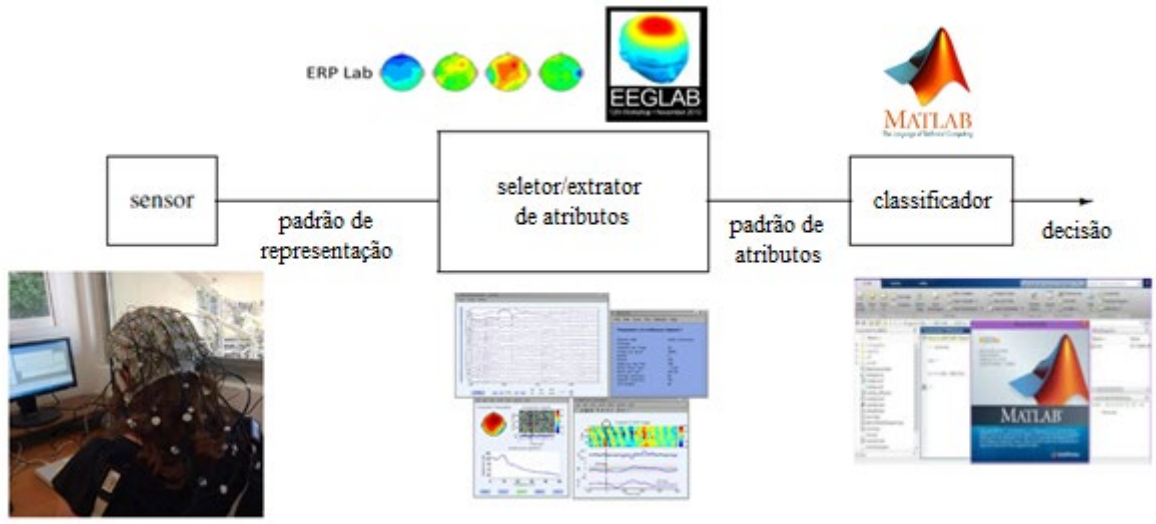
C. Softwares de extração de dados EEG / ERP e plataforma Matlab®

Para extrair e organizar os dados do ERP de um experimento, existem vários softwares que auxiliam nessa atividade de mineração de dados. Com relação a este trabalho, foram utilizados o EEGLAB® e o ERPLAB®, que são as caixas de ferramentas Matlab® para processamento e análise de dados de EEG e ERP. Para o processo digital e estudo de reconhecimento de padrões, o software Matlab® foi utilizado.

D. Teoria de Reconhecimento de Padrões

Os sistemas de reconhecimento de padrões são, em muitos casos, treinados a partir de dados com suas classes previamente conhecidas (aprendizado supervisionado ou discriminação). Mas quando não há dados com as classes já rotuladas disponíveis, outros algoritmos podem ser usados para descobrir padrões previamente desconhecidos (aprendizado não supervisionado ou agrupamento). Webb (2012) define que, na classificação supervisionada, um conjunto de amostras de dados (cada uma consistindo de medidas em um conjunto de variáveis ou atributos que podem ser extraídos) está associado com o que corresponde aos tipos de classe. Essas classes e atributos são usados no projeto do classificador. Na classificação não supervisionada, os rótulos de dados (classes) não são conhecidos e é necessário buscar por grupos nos dados com as mesmas características que possam distinguir um grupo (classe) de outros. Conforme descrito por Webb (2012), um procedimento simplificado de reconhecimento de padrões é mostrado na Figura 5 com as funções de todas as origens experimentais de dados e ferramentas de software usadas neste trabalho.

Figura 5 - Método de Reconhecimento de Padrões e suas funções com as ferramentas de software usadas neste trabalho (Adaptado de Webb (2012))

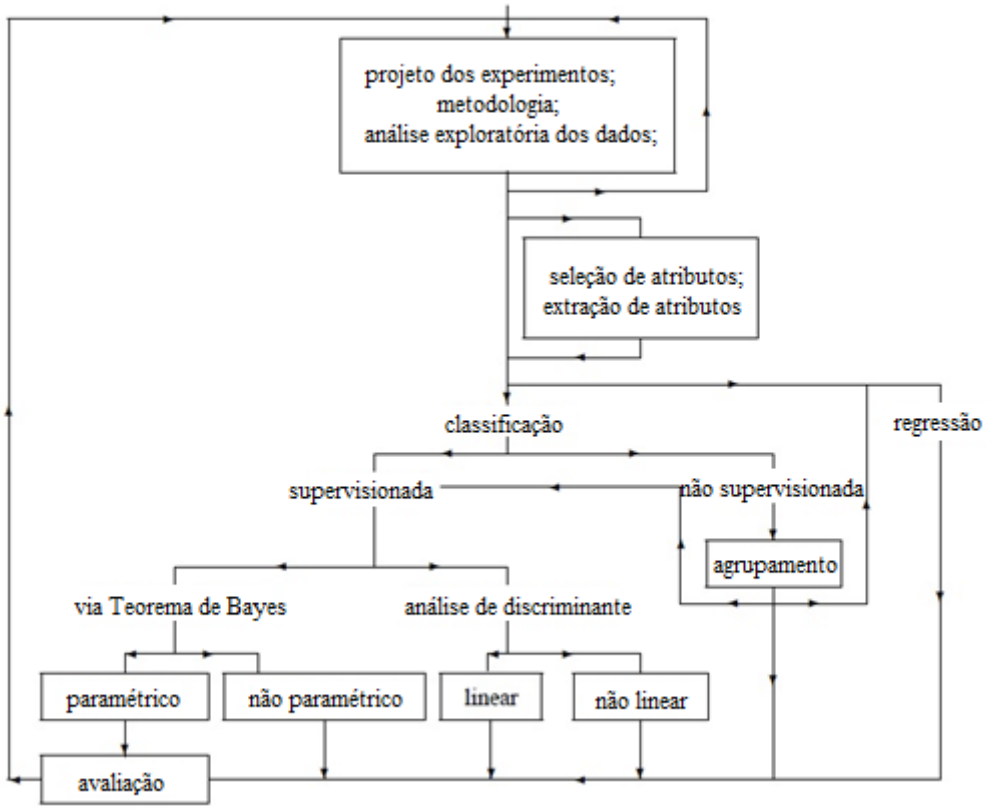


Como mostrado na Figura 5, neste estudo, o experimento de Soto (2014) está relacionado ao bloco Sensor (na verdade, o experimento EEG / ERP) e o seletor / extrator de recursos está relacionado com os softwares EEGLAB® e ERPLAB®. O bloco do classificador foi elaborado com o software Matlab®.

III. METODOLOGIA

A metodologia considerada neste trabalho baseia-se nas etapas de um problema de reconhecimento de padrões indicado por Webb (2012). A Figura 6 mostra esse procedimento de uma forma de fluxograma:

Figura 6 - Metodologia utilizada de Reconhecimento de Padrões proposta por Webb (2012)



De fato, a metodologia adotada utiliza as seguintes etapas de Webb (2012) adaptadas:

1. Formulação do problema, coleta de dados e exame inicial dos dados;
2. Seleção de recursos ou extração de recursos;
3. Classificação de padrões não supervisionados ou agrupamento;
4. Classificação de padrões supervisionados;
5. Avaliação dos resultados e Interpretação

Conforme mencionado no Resumo, este artigo foca nos resultados da classificação de padrões supervisionados, uma vez que os resultados dos classificadores não supervisionados obtidos foram discutidos em artigo anterior (ISSMAEL Jr., GESUALDI e CALVANO, 2017). Neste estudo do experimento Soto (2012), conforme já indicado anteriormente, os atributos que foram extraídos por meio dos softwares EEGLAB® e ERPLAB® para as tarefas de palavras e sentenças são os parâmetros do sinal ERP amplitude média entre duas latências fixas, amplitude de pico, latência de pico. Além desses atributos, também foram incluídos o intervalo de tempo do sinal ERP, a região de interesse (ROI) de onde foi extraído o sinal, e o sujeito humano relacionado a cada medição.

Em relação às classes consideradas para os classificadores, para a Tarefa "Sentenças" são: S1 (CSC), S2 (CNSC), S3 (ISC), S4 (INSC) e S5 (Controle) e, para a Tarefa "Palavras" são: S1 (SSR), S2 (ASR), S3 (Controle 1 - UR) e S4 (Controle 2 - PW).

Para fins de classificação, antes de construir o classificador, é necessário utilizar valores numéricos únicos para os atributos e classes, para permitir a convergência dos métodos de classificação do Matlab®. Atributos ou classes que não são variáveis numéricas como, por exemplo, a Região de Interesse (ROI), devem ser codificados com valores numéricos com uma correspondência coerente com o valor original da *"string"*.

Nas Tabelas I e II abaixo, são apresentadas a organização dos dados e codificação de valores para os atributos e classes para a elaboração dos classificadores.

TABELA I. ORGANIZAÇÃO E CODIFICAÇÃO PARA OS ATRIBUTOS DOS CLASSIFICADORES

Atributo	Valor	Código para o algoritmo do Matlab®
Faixa de Tempo do sinal ERP	150-300ms	1
	300-500ms	2
	500-700ms	3
Região de Interesse (ROI)	Linha Média Frontal	1
	Linha Média Central	2
	Linha Média Parietal	3
	Linha Média Occipital	4
	Lado Esquerdo Frontal	5
	Lado Esquerdo Central	6
	Lado Esquerdo Parietal	7
	Lado Esquerdo Occipital	8
	Lado Direito Frontal	9
	Lado Direito Central	10
	Lado Direito Parietal	11
	Lado Direito Occipital	12
Sujeito	2	2
	3	3
	4	4

	5	5
	6	6
	7	7
	8	8
	9	9
	10	10
	13	13
	15	15
	16	16
	17	17
	18	18
	19	19
	20	20

TABELA II. ORGANIZAÇÃO E CODIFICAÇÃO PARA AS CLASSES

Tarefa	Classes	Código para o algoritmo do Matlab®
Palavras	S1 (SSR)	1
	S2 (ASR)	2
	S3 (Controle 1 - UR)	3
	S4 (Controle 2 - PW)	4
Sentenças	S1 (CSC)	1
	S2 (CNSC)	2
	S3 (ISC)	3
	S4 (INSC)	4
	S5 (Control)	5

Após o uso do EEGLAB® e do ERPLAB®, os seguintes dados de entrada para os classificadores foram extraídos:

a) Para a Tarefa "Sentenças": uma matriz com 2880 linhas e 7 colunas, onde os atributos são correspondentes a: Coluna A: Amplitude Média entre duas latências fixas; Coluna B: Amplitude de Pico; Coluna C: Latência de pico; Coluna D: Região de Interesse (ROI); Coluna

E: Faixa de Tempo do Sinal ERP; Coluna F: índice do sujeito. A Coluna G corresponde as Classes das medidas.

b) Para a Tarefa "Palavras": uma matriz com 2304 linhas e 7 colunas, onde os atributos são organizados de forma similar à Tarefa "Sentenças".

A Figura 7 mostra essa organização de dados de entrada.

Figura 7 - Organização dos dados de entrada extraídos do EEGLAB® e do ERPLAB® para os classificadores.

	A	B	C	D	E	F	G
1	-2,049	2,227	254	1	1	2	1
2	-1,794	-0,304	196	1	1	3	1
3	1,099	4,893	260	1	1	4	1
4	-0,339	2,548	294	1	1	5	1
5	-2,309	3,372	266	1	1	6	1
6	-0,589	2,96	180	1	1	7	1
7	1,438	5,267	206	1	1	9	1
8	3,277	9,979	274	1	1	10	1
9	1,645	4,17	200	1	1	13	1
10	-0,572	1,157	260	1	1	15	1
⋮				⋮			⋮

Coluna A: Amplitude Média entre duas latências fixas;
Coluna B: Amplitude de Pico;
Coluna C: Latência de pico;
Coluna D: Região de Interesse (ROI);
Coluna E: Faixa de Tempo do Sinal ERP;
Coluna F: índice do sujeito; e
Coluna G: Classes.

Para este estudo da classificação supervisionada, os conjuntos de dados de Palavras e Sentenças foram divididos em três conjuntos com a mesma quantidade de dados com todos os atributos para cada classe proveniente de cada Tarefa. Os conjuntos são definidos como conjunto de treinamento, conjunto de validação e conjunto de teste, respectivamente, com 1/3 do total de dados existentes. Os métodos supervisionados do Matlab® usados para criar os scripts de algoritmos dos classificadores são o "Naïve Bayes", a Máquina de Vetor de Suporte (SVM) Multiclasse (SVM), Rede Neural e "Random Forest".

A figura de mérito usada para obter os resultados para os classificadores supervisionados foi a acurácia que é o número de previsões corretas de todas as previsões feitas. Para cada um dos métodos de classificação utilizados, foram efetuadas alterações manuais de parâmetros próprios de cada classificador, até se conseguir atingir a melhor acurácia.

IV. RESULTADOS E DISCUSSÃO

Os resultados serão apresentados considerando a melhor Acurãcia Total alcançada, para cada classificador. Para os classificadores supervisionados, os melhores resultados para a tarefa Sentenças são apresentados na Tabela III.

TABELA III. RESULTADOS PARA OS CLASSIFICADORES NA TAREFA SENTENÇAS

Classificador	Acurácias	Parâmetros do classificador utilizado, conforme nomenclatura do Matlab®
Naïve Bayes	33,7 %	Distribuiton function “kernel”
	97,3 %	Distribuiton function “MVMN”
Máquina de Vetor de Suporte Multiclasse	99,9 %	“BoxConstraint”:0.01 “KernelFunction”:“Gaussian” “Standardize”:“off”’
Rede Neural	27,2 %	a) Number of hidden layers (“hiddenLayerSize”): 10 b) Neural Network Input Processing Function (“net.input.processFcns”): 'removeconstantrows','mapstd' c) Neural Network Output Processing Function (“net.output.processFcns”): 'removeconstantrows','mapstd' d) Setup Division of Data for Training, Validation, Testing (“net.divideFcn“): 'dividerand' e) Train Ratio ("net.divideParam.trainRatio") = 1/3 f) Validation Ratio ("net.divideParam.valRatio") = 1/3; g) Test Ratio ("net.divideParam.testRatio") = 1/3 h) Divide Mode (“net.divideMode”): 'sample' i) Multilayer Neural Network Training Function (“net.trainFcn”):'trainrp'

		j) Neural Network Performance Function (“net.performFcn”): 'mse'
Random Forest	100,0 %	a) 'method' (Ensemble-aggregation method): “bag” b) 'NLearn' (Number of ensemble learning cycles): 100 c) 'Learners' (Weak learners to use in ensemble): “Tree” d) 'Type' (Supervised learning type): 'classification'

Os melhores resultados para a Tarefa Palavras são apresentados na Tabela IV.

TABELA IV. RESULTADOS PARA OS CLASSIFICADORES NA TAREFA PALAVRAS

Classificador	Acurácias	Parâmetros do classificador utilizado, conforme nomenclatura do Matlab®
Naïve Bayes	39,2 %	Distribuiton function “kernel”
	97,6 %	Distribuiton function “MVMN”
Máquina de Vetor de Suporte Multiclasse	99,9 %	“BoxConstraint”: 0.01 “KernelFunction”: “Gaussian” “Standardize”: “off”’
Rede Neural	35,2 %	a) Number of hidden layers (“hiddenLayerSize”): 60 b) Neural Network Input Processing Function(“net.input.processFcns”): 'removeconstantrows','mapstd' c) Neural Network Output Processing Function(“net.output.processFcns”): 'removeconstantrows','mapstd' d) Setup Division of Data for Training,

		Validation, Testing (“net.divideFcn“): 'dividerand'
		e) Train Ratio ("net.divideParam.trainRatio") = 1/3;
		f) Validation Ratio ("net.divideParam.valRatio") = 1/3;
		g) Test Ratio ("net.divideParam.testRatio") = 1/3;
		h) Divide Mode (“net.divideMode”): 'sample';
		i) Multilayer Neural Network Training Function (“net.trainFcn”):'trainscg'
		j) Neural Network Performance Function (“net.performFcn”):'mse'

Random Forest	100,0 %	a) 'method' (Ensemble-aggregation method): “bag” b) 'NLearn' (Number of ensemble learning cycles): 100 c) 'Learners' (Weak learners to use in ensemble): “Tree” d) 'Type' (Supervised learning type): 'classification'

A campanha de teste considerou os conjuntos de dados completos para ambas as Tarefas. Ao final da metodologia proposta, com o intuito de aprofundar a investigação das características utilizadas na classificação, especificamente sobre os sujeitos (as pessoas submetidas ao experimento), foram realizados testes de classificação sem se fazer a divisão do conjunto de dados em treinamento e teste, para todos os classificadores supervisionados, para ambas as tarefas, sem reciclagem de dados, e executando os algoritmos individualmente para cada sujeito.

Como o comportamento do cérebro entre os indivíduos pode ser bem diferente, embora, os perfis das pessoas que participaram dos experimentos sejam semelhantes (SOTO, 2014), o objetivo é verificar se os resultados para os indivíduos podem ser diferentes em relação aos conjuntos de dados completos para todos eles. Os resultados podem ser vistos na Tabela V, para a Tarefa “Sentenças”, e na Tabela VI, para a Tarefa “Palavras”.

TABELA V. RESULTADOS PARA OS CLASSIFICADORES PARA INDIVÍDUOS NA TAREFA "SENTENÇAS"

Aplicando Discriminação (Classificadores Supervisionados)						Método de Regressão	
Naïve Bayes MVMN		SVM		Rede Neural		Random Forest	
Sujeito	Acurácia %	Sujeito	Acurácia %	Sujeito	Acurácia %	Sujeito	Acurácia %
2	100.00%	2	100.00%	2	87,80%	2	100.00%
3	98.89%	3	99.44%	3	23,90%	3	100.00%
4	100.00%	4	99.44%	4	32,20%	4	100.00%
5	100.00%	5	100.00%	5	52,80%	5	100.00%
6	100.00%	6	100.00%	6	34,40%	6	100.00%
7	100.00%	7	100.00%	7	20,00%	7	100.00%
9	99.44%	9	100.00%	9	44,40%	9	100.00%
10	100.00%	10	100.00%	10	23,30%	10	100.00%
13	100.00%	13	100.00%	13	33,90%	13	100.00%
15	100.00%	15	100.00%	15	19,40%	15	99.44%
16	100.00%	16	100.00%	16	45,00%	16	100.00%
17	99.44%	17	100.00%	17	25,60%	17	100.00%
18	100.00%	18	100.00%	18	43,30%	18	100.00%
19	100.00%	19	100.00%	19	61,70%	19	100.00%
20	100.00%	20	100.00%	20	36,10%	20	100.00%
21	100.00%	21	100.00%	21	29,40%	21	100.00%
Acurácia do classificador com conjunto de dados completo:		Acurácia do classificador com conjunto de dados completo:		Acurácia do classificador com conjunto de dados completo:		Acurácia do classificador com conjunto de dados completo:	

97.26%	99.90%	27,20%	100.00%

TABELA VI. RESULTADOS PARA OS CLASSIFICADORES PARA INDIVÍDUOS NA TAREFA "PALAVRAS"

Aplicando Discriminação (Classificadores Supervisionados)						Método de Regressão	
Naïve Bayes MVMN		SVM		Naïve Bayes MVMN		SVM	
Sujeito	Acurácia %	Sujeito	Acurácia %	Sujeito	Acurácia %	Sujeito	Acurácia %
2	100.00%	2	100.00%	2	26,40%	2	100.00%
3	100.00%	3	99.31%	3	34,70%	3	100.00%
4	100.00%	4	99.31%	4	37,50%	4	100.00%
5	100.00%	5	100.00%	5	25,70%	5	100.00%
6	99.31%	6	100.00%	6	34,70%	6	100.00%
7	100.00%	7	100.00%	7	30,60%	7	100.00%
9	100.00%	9	100.00%	9	50,00%	9	100.00%
10	100.00%	10	100.00%	10	28,50%	10	100.00%
13	100.00%	13	99.31%	13	36,10%	13	100.00%
15	100.00%	15	100.00%	15	25,00%	15	100.00%
16	100.00%	16	100.00%	16	27,80%	16	100.00%
17	100.00%	17	100.00%	17	34,70%	17	100.00%
18	100.00%	18	100.00%	18	52,10%	18	100.00%
19	100.00%	19	100.00%	19	32,60%	19	100.00%
20	100.00%	20	100.00%	20	41,00%	20	100.00%
21	100.00%	21	100.00%	21	21,50%	21	100.00%

Acurácia do classificador com conjunto de dados completo: 97,60%	Acurácia do classificador com conjunto de dados completo: 99,90%	Acurácia do classificador com conjunto de dados completo: 35,20%	Acurácia do classificador com conjunto de dados completo: 100.00%

Os resultados indicam que todos os indivíduos, de forma independente, atingiram 100% de acurácia para o classificador "Random Forest", com exceção de um indivíduo (sujeito 15), na Tarefa "Sentenças", que atingiu 99,44%. Para os classificadores "Naïve Bayes" e "SVM" multiclasses alguns poucos sujeitos apresentaram diferentes acurácias, mas, mesmo assim, com valores superiores a 98%. O pior caso foi o classificador da Rede Neural, para o qual todos os indivíduos apresentaram diferentes acurácias, indicando insucesso na classificação.

V. CONCLUSÕES

O objetivo deste trabalho, que foi investigar a metodologia de reconhecimento de padrões de Webb (2012) em resultados de sinais ERP do experimento linguístico de Soto (2014), para se classificar corretamente diferentes padrões, foi considerado como atingido.
As ferramentas de software EEGLAB®, ERPLAB® e Matlab® utilizadas para executar etapas de pré-processamento e reconhecimento de padrões nos dados de Soto (2014) funcionaram a contento, facilitando o trabalho de organização e extração de atributos e classes para os classificadores.

Como mostrado neste artigo, as abordagens de classificadores supervisionados não-lineares atingiram acurácias superiores a 96%, sendo assim consideradas como mais adequadas para este tipo de campanha de classificação. O cenário de classificação que utilizou o método de classificação supervisionado "Random Forest" obteve melhor resultado para esses conjuntos de dados, com uma acurácia total de 100%. Outros bons resultados obtidos utilizaram os classificadores supervisionados SVM multiclasse e "Naïve Bayes", com acurácia total superior a 96%.

Estes resultados permitem também concluir que para estes conjuntos de dados ERP, tanto para a Tarefa "Sentenças", quanto para a Tarefas "Palavras", as abordagens não lineares foram mais adequadas para classificar os dados da configuração experimental de Soto (2014). Este resultado também se mostrou válido para cada sujeito e grupo de sujeitos.

Outros métodos de classificação e, especialmente, para agrupamento e classificação não supervisionada (ISSMAEL Jr., GESUALDI e CALVANO, 2017). devem ser considerados para estudar diferentes aspectos deste conjunto de dados para reconhecimento de padrões, visando a se apresentar como um método importante e, salvo melhor juízo, inédito para diagnósticos em neurolinguística e medicina, através da identificação de quais atributos do sinal ERP podem ser mais influentes no processo de classificação.

Mesmo com esses bons resultados, o próximo passo é dar continuidade aos estudos em relação à análise dos classificadores propostos, ampliando a verificação em relação aos demais atributos, especialmente as ROI e as faixas de tempo do sinal ERP.

AGRADECIMENTOS

Os autores gostariam de agradecer a Dra. Marije Soto e ao Centro Federal de Tecnologia Celso Suckow da Fonseca por permitirem a execução deste trabalho.

REFERÊNCIAS

Soto, Marije, ERP and fMRI Evidence of Compositional Differences between Linguistic Computations for Words and Sentences. Marije Soto - Rio de Janeiro:UFRJ./Faculdade de Letras, 2014.

Webb, Andrew R., Statistical Pattern Recognition, Second Edition. John Wiley & Sons, Ltd. 2012. ISBNs: 0-470-84513-9 (HB); 0-470-84514-7 (PB)

Gesualdi, Aline da Rocha; França, Aniela Improta. Event-related brain potentials (ERP): an overview. Revista Linguística / Revista do Programa de Pós-Graduação em Linguística da Universidade Federal do Rio de Janeiro. Volume 7, número 2, dezembro de 2011. ISSN 1808-835X 1. [http://www.letras.ufrj.br/poslinguistica/revistalinguistica]

Luck, Steven J., An Introduction to the Event-Related Potential Technique, Massachusetts Institute of Technology MIT Press books, 2nd Ed., 2014, ISBN 978-0-262-52585-5

Woodman, Geoffrey F.. A Brief Introduction to the Use of Event-Related Potentials (ERPs) in Studies of Perception and Attention. Atten Percept Psychophys. 2010 November; 72(8): doi:10.3758/APP.72.8.2031.

Eeglab®. EEGLAB® Tutorial. Available on: <http://sccn.ucsd.edu/wiki/Getting_Started>, accessed in April, 15th, 2016.

Erplab®. ERPInfo-ERPLAB® Toolbox. Available on: <http://www.erpinfo.org/erplab.html>, consulted on April, 15th, 2016.

Lopez-Calderón, Javier, Luck, Styeven J. ERPLAB: an open-source toolbox for the analysis of event-related potentials. Frontiers in Human Neuroscience. Volume 8. Article 213. p. 2-14. April 2014.

Matlab®. MATLAB® - The Language of Technical Computing. Available on: <https://www.mathworks.com/products/matlab.html>, consulted on April, 15th, 2016a.

Issmael Jr., A.K., Gesualdi Manhães, Aline, Calvano, José Vicente. Application of Pattern Recognition Method in a Linguistic Experiment with Unsupervised Classification. Anais do XXXV Simpósio Brasileiro de Telecomunicações e Processamento de Sinais - SBrT 2017. Sao Pedro. Sao Paulo. Brazil. p.905-909. 03-06 sep 2017.