

**Universidade de São Paulo
Faculdade de Medicina de Ribeirão Preto
Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto**

MINERAÇÃO DE DADOS APLICADA A EVENTOS
NEUROFISIOLÓGICOS ATRAVÉS DE EEG USANDO
MODELOS DE EPILEPSIA EXPERIMENTAL

Diego de Caiado Castro Borragini

Ribeirão Preto
2008

MINERAÇÃO DE DADOS APLICADA A EVENTOS
NEUROFISIOLÓGICOS ATRAVÉS DE EEG USANDO
MODELOS DE EPILEPSIA EXPERIMENTAL

Diego de Caiado Castro Borragini

Orientador: Prof. Dr. Norberto Garcia-Cairasco

Co-Orientador: Prof. Dr. José Augusto Baranauskas

Monografia apresentada à Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto e à Faculdade de Medicina de Ribeirão Preto como requisito para a obtenção do título de Bacharel em Informática Biomédica.

Diego de Caiado Castro Borragini

**MINERAÇÃO DE DADOS APLICADA A EVENTOS
NEUROFISIOLÓGICOS ATRAVÉS DE EEG USANDO
MODELOS DE EPILEPSIA EXPERIMENTAL**

Monografia apresentada à Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto e à Faculdade de Medicina de Ribeirão Preto como requisito para a obtenção do título de Bacharel em Informática Biomédica.

Orientador: Prof. Dr. Norberto Garcia-Cairasco
Co-Orientador: Prof. Dr. José Augusto Baranauskas

COMISSÃO EXAMINADORA

Prof. Dr. Norberto Garcia-Cairasco
Faculdade de Medicina de Ribeirão Preto

Prof. Dr. Cléver Ricardo Guareis de Farias
Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto

Prof. Dr. Renato Tinós
Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto

Ribeirão Preto, 15 de Dezembro de 2008.

Agradecimentos

Agradeço a todos os colegas do LNNE pela acolhida no meio científico e principalmente pela paciência em transmitir todo o conhecimento que adquiri ao longo dos três últimos anos.

Em especial agradeço a meu Professor e Orientador Norberto Garcia-Cairasco não só pelos ensinamentos mas também por ser um grande incentivador da abordagem multidisciplinar da biologia, o que proporcionou a realização desse trabalho e minha iniciação na Neurociências.

Agradeço a Marcelo Cairrão por ser um dos grandes responsáveis pelos primeiros passos de minha iniciação científica. E também, a Artur Fernandes, Maira Foresti e Flávio Del Vecchio não só por fornecer material para a realização desse trabalho mas também por todas as explicações e seções de conversas que tivemos.

Agradeço a meu Professor e Co-Orientador José Augusto Baranauskas por acreditar que valia a pena “bater na mesma tecla” e fazer com que eu me interessasse pelo que é hoje parte do conhecimento utilizado nesse trabalho.

Agradeço a todos os amigos e colegas que me acompanharam durante esses anos e contribuíram com apoio, discussões, ensinamentos e principalmente com a amizade. Em especial, a Gustavo Rocha e André Villa por serem como irmãos e lembrarem sempre do “Todo mundo junto e misturado”.

Agradeço à minha família pela força e apoio em todos os momentos de minha vida e principalmente à minha mãe, Marília, por ser quem acompanhou diariamente minha trajetória ao longo da faculdade.

Em especial, dedico o sucesso que tive durante esses anos em memória de meu pai, Moisés, que foi um verdadeiro professor e também a primeira pessoa a dizer que sentia orgulho por seu filho buscar o que quer.

Agradeço à FAPESP, FAPESP-Cinapce, CNPq, CAPES e FAEPA pelo apoio às pesquisas do LNNE.

Resumo

Esse trabalho tem como objetivo apresentar um estudo sobre a utilização de técnicas matemáticas e computacionais para o processamento de sinais de eletrencefalografia (EEG) oriundos de modelos de epilepsia experimental.

Através do processamento de sinais são realizadas a aplicação e a comparação de métodos para extração de informações relevantes do traçado do EEG de ratos através do reconhecimento de padrões por mineração de dados. Tais informações são utilizadas em algoritmos computacionais de indução, visando prever novos eventos neurofisiológicos como, por exemplo, a entrada no período ictal, ou seja, em uma crise epiléptica e também a saída dessa com o retorno ao período basal do animal.

O estudo mostrou que é possível tal predição por algoritmos de mineração de dados utilizando informações intrínsecas no sinal de EEG. A eficácia da metodologia de aprendizagem e predição dos eventos de interesse girou em torno de 93%-97% dos casos de análise, sugerindo que ainda é possível observar novos dados ou mesmo realizar um processamento mais detalhado das informações a fim de se melhorar o desempenho da metodologia.

Palavras Chaves: Eletrencefalografia, Mineração de Dados, Epilepsia, Transformada Fourier

Abstract

This project aims at presenting the use of computational and mathematical techniques to process electroencephalography (EEG) signals from experimental models of epilepsy.

The implementation and comparison of methods to extract information by recognizing patterns in data mining process is done through signal processing. Such information with computer algorithms is capable to predict new neurophysiological events, such as the beginning of the ictal period, i.e. the onset of an epileptic seizure and also its end-point of that, with the return to the baseline of the animal behavior.

The study indicates that this prediction is possible with algorithms of data mining, by means of information inherent to the EEG signal. The effectiveness of the methodology of learning and prediction of events of interest solved about 93% -97% of analyzed cases, suggesting that it is still possible to see new data or even conduct a more detailed processing of information in order to improve performance of the methodology.

Key words: Electroencephalography, Data Mining, Epilepsy, Fourier Transform

Sumário

CAPÍTULO 1	1
1.1 INTRODUÇÃO	1
1.2 OBJETIVOS	6
1.3 ORGANIZAÇÃO DO DOCUMENTO	7
CAPÍTULO 2	8
2.1 CONCEITOS MATEMÁTICOS	8
2.1.1 ANÁLISE ESPECTRAL	8
2.2 CONCEITOS COMPUTACIONAIS	9
2.2.1 MINERAÇÃO DE DADOS	12
CAPÍTULO 3 – METODOLOGIA E PROCEDIMENTOS	14
3.1 OBTENÇÃO DOS DADOS	14
3.1.1 ESTIMULAÇÃO	14
3.1.2 CAPTURA	15
3.1.3 REGISTRO	16
3.2 PRÉ-PROCESSAMENTO	17
3.2.1 ANÁLISE DE VÍDEO-EEG	17
3.2.2 CLASSIFICAÇÃO DE ÉPOCAS	18
3.2.3 EXTRAÇÃO DE INFORMAÇÕES	21
3.2.4 ARQUIVO ARFF	23
3.2.5 MINERAÇÃO DE DADOS	26
CAPÍTULO 4 – RESULTADOS	29
4.1 ANÁLISE COMPUTACIONAL	29
4.1.1 ESCOLHA DE PARÂMETROS	29
4.1.2 ALGORITMO J48	30
4.1.3 ALGORITMO PART	36
4.1.4 CONSIDERAÇÕES FINAIS	37
4.1.1 VALIDAÇÃO DA METODOLOGIA	39
CAPÍTULO 5 – CONCLUSÃO	42
BIBLIOGRAFIA	43
ANEXOS	46
ANEXO 1	47
ANEXO 2	48
ANEXO 3	49
ANEXO 4	50
ANEXO 5	51
ANEXO 6	52
ANEXO 7	53

Lista de Figuras

Figura 01 –	Etapas do KDD segundo Fayaad (Fayyad et al., 1996).....	11
Figura 02 –	Seleção dos registros disponíveis para início do estudo.....	18
Figura 03 –	Esquema ilustrativo do traçado EEGráfico.	20
Figura 04 –	Resultado do Algoritmo para Separar Épocas.	20
Figura 05 –	Diagrama de Sinal Simulado: Análise Espectral por Fourier.....	21
Figura 06 –	Diagrama de Sinal Simulado: Análise Espectral por Wavelet	22
Figura 07 –	Exemplo de arquivo ARFF para uso na ferramenta Weka.	24
Figura 08 –	Exemplo de Árvore de Decisão. (Waikato 2004).....	27

Lista de Tabelas

Tabela 01 – Exemplo de Matriz de Confusão na Classificação de Exemplos.....	30
Tabela 02 – Classificações corretas do J48, em porcentagem – minNumObj=1	31
Tabela 03 – Classificações corretas do J48, em porcentagem – minNumObj=2	32
Tabela 04 – Classificações corretas do J48, em porcentagem – minNumObj=3	32
Tabela 05 – Melhor Desempenho de acordo com o J48 – minNumObj	33
Tabela 06 – Representação dos Índices <i>Kappa</i> na classificação do J48.....	33
Tabela 07 – Melhores Índices <i>Kappa</i> na Classificação do J48	34
Tabela 08 – Representação da Sensibilidade e da Especificidade do J48.....	35
Tabela 09 – Desempenho do PART com os Conjuntos 06 e 07 – Classificação Correta.....	36
Tabela 10 – Desempenho do PART com os Conjuntos 06 e 07 – F-Measure	37

Capítulo 1

1.1 Introdução

Em trabalhos pioneiros do final do século XIX e início do século XX, Julius Bernstein fez a primeira descrição apurada do potencial de ação, bem como publicou sua teoria, em parte válida até hoje, sobre potenciais elétricos em membranas biológicas (Julius Bernstein apud (Seyfarth, 2006). No entanto, a visualização temporal contínua das variações de voltagem produzidas no encéfalo (eletrencefalograma, EEG), registrada na calota craniana intacta de seres humanos só foi efetivamente desenvolvida em 1929 por Hans Berger (Berger, 1929).

A técnica de eletrencefalografia possibilita o registro das atividades elétricas normais e patológicas ocorridas no cérebro, quer seja durante fenômenos fisiológicos como sono, vigília alerta e vigília relaxada, ou durante quadros patológicos como crises epiléticas. Essas atividades são registradas a partir de oscilações de voltagem e ritmos variados de frequências que ocorrem entre áreas cerebrais, se alterando de acordo com o estado comportamental do paciente (Paré et al., 2002).

No caso de eventos fisiológicos, é possível observar padrões EEGráficos característicos como, por exemplo, ritmos de frequências denominados “beta” (14 Hz a 30 Hz) em vigília alerta, ritmos “alfa” (8 Hz a 13 Hz) em vigília relaxada e predomínio da faixa “teta” (0,5 Hz a 7 Hz) durante a fase de sono sincronizado (Steriade, 1993).

Com relação às epilepsias, os padrões podem variar conforme o tipo de crise apresentada. Por exemplo, nas crises de ausência, praticamente sem manifestações motoras e mais comuns em crianças, é patente a presença de descargas tipo ponta-onda (*spike and wave*

discharges), envolvendo principalmente o circuito talâmico-cortical, na frequência de 3 Hz para humanos (JASPER and Droogleever-Fortuyn, 1949) e 7-8 Hz em modelos animais (Vergnes et al., 1987). As crises de ausência são classificadas como crises primariamente generalizadas, pois seu início ocorre abrupta e simultaneamente em ambos os hemisférios cerebrais (Niedermeyer et al., 1969), embora dados mais recentes indiquem que exista um foco epilético que se espalha muito rapidamente (Meeren et al., 2002). Por outro lado, durante crises convulsivas, a assinatura eletrencefalográfica é a presença de deflexões súbitas no registro, chamadas de espículas (*spikes*) (HOUDART et al., 1952;Engel, Jr. et al., 1975).

Além desses elementos característicos, sabe-se que há padrões EEGráficos no início das crises (*seizure onsets*) como, por exemplo, o *Low Voltage Fast – LVF* (Baixa Voltagem e aumento rápido de Frequência), e o Hipersincrônico (HIP, Espículas de Alta Amplitude, com Frequência menor que 2 Hz e duração maior que 5 s) ao iniciar ou logo após o início em crises recorrentes tanto em pacientes epiléticos como modelos animais de epilepsia (Bragin et al., 2005a;Bragin et al., 2005b). Já na fase final das crises, podemos observar a atenuação do traçado EEGráfico de modo proporcional à duração da crise (Sakamoto et al., 2000). Estes elementos EEGráficos (espículas, ondas lentas e os complexos ponta-onda) são bem característicos e destacam-se da atividade cerebral de fundo, mesmo apenas com inspeção visual dos registros. De acordo com esses elementos no traçado do EEG, podemos tanto diagnosticar epilepsias como também classificá-las como focais ou generalizadas, bem como estudar compostos convulsivantes ou anti-convulsivantes.

Dentre os vários fatores que podem alterar o estado fisiológico do cérebro e levar a crises epiléticas, encontramos os estímulos fóticos (Kasteleijn-Nolst Trenite, 1989;Parra et al., 2005), acústicos (Garcia-Cairasco et al., 1996), químicos (Ben Ari, 1985) e elétricos (Goddard, 1967;Racine et al., 1972;McNamara, 1984;Morimoto et al., 2004;Schubert et al., 2005).

O abrasamento (*kindling*) elétrico é um dos modelos experimentais mais utilizados no estudo da epilepsia, sendo realizado através de estimulação elétrica do cérebro do rato, isto é, o animal é submetido a estímulos elétricos repetitivos, dos quais resulta um desenvolvimento progressivo de crises comportamentais e eletrencefalográficas (Goddard, 1967).

Para o estudo dessas e outras alterações fisiológicas no cérebro utilizamos o traçado EEGráfico, o qual como já mencionado, pode ser analisado morfológicamente por meio da inspeção visual. Este método, muito empregado atualmente na clínica de distúrbios epilépticos e do sono, requer muito treinamento e está baseado apenas no domínio Amplitude de Voltagem vs. Tempo. Porém, através de ferramentas matemáticas chamadas coletivamente de *análise espectral* e com auxílio de técnicas computacionais, podemos avaliar o conteúdo de freqüência de oscilação nos registros, o qual se encontra intrínseco nos dados.

O surgimento desses novos métodos matemáticos para análise de sinais possibilitou, portanto, a extração de domínios espectrais dos períodos desejados de um EEG. Através desses domínios se torna possível o estudo dos fenômenos neurofisiológicos por análise de oscilações de voltagem, amplitudes, freqüências, entre outros dados que podem indicar um evento em particular.

No Laboratório de Neurofisiologia e Neuroetologia Experimental (LNNE), onde se estudam diversos modelos experimentais de epilepsia, bem como vídeos e registros de pacientes epilépticos do Centro de Cirurgia de Epilepsias do HC-FMRP-USP, há o interesse constante na possibilidade de análises matemáticas dos registros EEGráficos. Criou-se então um grupo interdisciplinar para estudo de *wavelets* e desenvolvimento de ferramentas matemáticas para análise de EEG em modelos experimentais de Epilepsia. Notoriamente o Prof. Dr. Dráulio Barros de Araújo¹ e o Dr. Rodrigo Neves Romcy Pereira² criaram

¹ Depto. Física e Matemática, FFCLRP-USP;

² Doutor pelo LNNE, Pós-Doutorado pela *Rockefeller University* e atualmente Bolsista PRODOC do Depto. Neurologia, Psiquiatria e Psicologia Médica, HC-FMRP-USP, sob supervisão do Prof. Dr. João Pereira Leite;

algoritmos para análise das principais frequências de voltagem que se sustentam ao longo do tempo (chamadas de frequências sustentadas, **FSS**). Esse trabalho resultou na apresentação de um pôster no *Joint International Meeting of the Physiological Society and the Brazilian Physiological Society*, 2006, Ribeirão Preto, e também na elaboração de um artigo já publicado (Romcy-Pereira et al., 2008).

Além do desenvolvimento de ferramentas matemáticas para o estudo de EEG, houve também progressos no estudo biológico utilizando tais ferramentas, inicialmente por meio da análise de *wavelets* (Rossetti et al., 2006) e, mais recentemente, na adaptação da metodologia descrita por Grinsted (Grinsted et al., 2004) para análise de coerência de sinais biológicos em modelos animais de epilepsia³. Evidencia assim, o interesse pela análise de dados experimentais através de técnicas matemáticas e computacionais para a extração de novas informações que auxiliam na busca pelo conhecimento sobre fenômenos neurofisiológicos.

Entretanto, apesar da importância do estudo matemático dos EEGs, do domínio espectral, da coerência entre áreas cerebrais e da existência de algoritmos para análise de coerências e de frequências sustentadas, não há atualmente em nosso laboratório um estudo capaz de constatar a presença de assinaturas EEGráficas em modelos de epilepsia experimental em animais de forma automática. A análise desses padrões epileptiformes permitiria ganho de informações importantes sobre o funcionamento dos núcleos encefálicos envolvidos em tais crises, assim como os mecanismos de ativação e desativação das crises nesses núcleos.

A proposta deste trabalho consiste na extração de informações relevantes do EEG e a busca por assinaturas epileptiformes na tentativa de se predizer um evento específico.

³ “Caracterização de Sistemas Anticonvulsivantes Endógenos”, Pós-Doutorado de Marcelo Cairrão Araújo Rodrigues – Processo FAPESP N° 03/01134-9

Contudo, o grande problema pela busca desses padrões no traçado EEGráfico é que a análise numérica desses registros se dá através de potenciais elétricos que são basicamente medidos em unidades de voltagem. Essas são tão numerosas por segundo quanto à resolução do equipamento de medição, o que tornaria inviável para uma simples análise visual humana e principalmente, de difícil distinção entre sinais de interesse e informações de fundo (ruído).

Para resolver esse impasse realizamos a aplicação da extração semi-automática de conhecimento a partir do grande volume de dados (KDD – *Knowledge Data Discovery*), definida por Fayyad (Fayyad et al., 1996) como: “Processo não trivial de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis nos dados existentes”.

Abordamos também a Mineração de Dados (*Data Mining*) que é a fase do KDD responsável pela seleção dos métodos a serem utilizados para localizar padrões nos dados disponíveis, pela busca efetiva dos padrões e pelo melhor ajuste nos parâmetros do algoritmo utilizado. Essa última fase geralmente é realizada utilizando-se algoritmos de Aprendizado de Máquina (AM), que é uma área de pesquisa em Ciência da Computação que tem crescido rapidamente nos últimos anos devido à existência de tecnologias poderosas para aquisição e gerenciamento de grande quantidade de dados.

O LNNE possui abundância de dados EEGráficos oriundos de modelos experimentais de epilepsia, que mesmo tendo rendido publicações internacionais, dissertações e teses, ainda guardam informações relevantes. O correto uso de ferramentas matemáticas para quantificação de parâmetros do EEG não apenas resultará em novas informações contidas nestes registros, mas também representará um uso ético e racional dos animais de experimentação, evitando que novos animais sejam implantados em experimentos futuros apenas para a coleta do EEG. Certamente que a validação destas metodologias nos modelos experimentais também justificará a implementação e uso futuro destas ferramentas em registros oriundos da clínica neurológica.

A partir de estudos anteriores (Rodrigues et al., 2005), verificamos que certos comportamentos ou seqüências desses se repetem concomitantemente COM OS traçados epileptiformes, como por exemplo, a realização do comportamento “*Wet Dog Shake*” (sacudidas de cachorro molhado) para se resgatar a função de movimentação normal do animal no período ictal, ou seja, durante a crise. Essas repetições juntamente com padrões de ritmos em fenômenos neurofisiológicos, como citado anteriormente, levantam a questão da presença de assinaturas EEGráficas, como por exemplo, no período pré-ictal (anterior à crise) indicando o início do evento de interesse. É interessante salientar que dependendo do modelo experimental estudado a presença de “*wet dog shakes*” pode sinalizar, ora eventos convulsivantes, ora eventos anti-convulsivantes (Rodrigues et al., 2005).

O significado dessas assinaturas, assim como a busca por esses padrões EEGráficos / Comportamentais, correlacionando o traçado epileptiforme, a oscilação de frequências e potenciais de ação, e o evento comportamental no qual se encontra o animal no período de análise, possibilita ganho de novos dados, antes invisíveis a análise morfológica, que auxiliam no entendimento dos mecanismos de ativação e desativação de crises epiléticas.

1.2 Objetivos

O objeto de estudo desse trabalho foi constituído por um grupo de registros eletrencefalográficos proveniente de animais submetidos ao modelo de Abrasamento elétrico rápido da Amígdala, conforme o protocolo experimental da Doutora Maira Licia Foresti (Foresti et al., 2008). Dois outros grupos de dados foram estudados, os provenientes do projeto de Iniciação Científica (I.C.) da aluna Érika Santos, com o modelo de *Status*

Epilepticus (SE) induzido pela injeção de pilocarpina intrahipocampal e o do Mestre Artur Fernandes (Fernandes, 2008), com o modelo de SE induzido pela estimulação elétrica contínua do complexo amigdalóide seguindo o protocolo desenvolvido em nosso laboratório por Cristiane Tilelli (Tilelli et al., 2005).

O objetivo geral desse trabalho foi, portanto, o estudo e adaptação da metodologia existente para análise de sinais eletrencefalográficos através de técnicas matemáticas e quantitativas, para a aplicação de algoritmos de Aprendizado de Máquina visando uma possível identificação de assinaturas EEGráficas de fenômenos neurofisiológicos, particularmente no modelo de epilepsia experimental de *kindling* rápido por estimulação elétrica da amígdala.

1.3 Organização do documento

O restante desse documento está dividido da seguinte forma: **Capítulo II** – Estado da Arte. Conceitos adicionais sobre a Matemática e a Computação utilizadas no decorrer desse trabalho. **Capítulo III** – Metodologia e Procedimentos adotados. **Capítulo IV** – Resultados Obtidos e Análises. **Capítulo V** – Conclusões.

Capítulo 2

Esse capítulo apresenta os conceitos relacionados às técnicas matemáticas e computacionais utilizadas no processamento de sinais e da etapa de mineração de dados deste trabalho.

2.1 Conceitos Matemáticos

2.1.1 Análise Espectral

A técnica matemática de Análise Espectral nos permite estudar o comportamento de um sinal qualquer através do domínio do Tempo *vs.* domínio da Frequência.

A análise mais comumente empregada é a Transformada de Fourier (FT - *Fourier Transform*). Desenvolvida no início do século XIX, por Jean-Baptiste Joseph Fourier (Fourier, 1822), a técnica consiste em dividir uma função contínua em componentes senóides que nos permite identificar as frequências de acordo com as amplitudes espectrais. Sua principal aplicação na época foi na elaboração da “Teoria de Condução do Calor”. Atualmente essa transformada é muito utilizada em diversas áreas das ciências exata, incluindo principalmente o processamento de sinais em telecomunicações, cujas frequências não variam com o tempo e são ditos sinais estacionários, e também auxiliam estudos na área biológica.

Entretanto, é necessário ressaltar que a teoria da Transformada de Fourier é utilizada com restrições nos estudos de sinais biológicos devido à ocorrência de oscilações de frequências no decorrer do tempo, uma vez que esses sinais são em sua maioria não-estacionários. Assim, se quisermos observar o sinal no domínio do Tempo *vs.* domínio da

Frequência utilizando a FT, devemos implementar a proposta de Gabor e aplicar um janelamento na FT, sendo assim conhecida por *Short Time Fourier Transform* (STFT) ou Transformada de Gabor (Kiyimik et al., 2005). Através desse janelamento conseguimos extrair o domínio Tempo vs. domínio Frequência exatamente no momento desejado.

Contudo, a STFT possui ainda a limitação do janelamento fixo, isto é, o tamanho da janela escolhida para análise espectral influenciará muito o resultado, pois ondas muito lentas precisam de uma janela temporal maior para se repetirem algumas vezes, em contrapartida ondas mais rápidas (com maior frequência) completam seus ciclos mais rapidamente e podem ser analisadas com uma janela menor. Para contornar o problema do janelamento fixo da STFT, podemos então realizar a Transformada de *Wavelet* (WT), a qual considera a faixa espectral do sinal localizada na janela temporal que podemos estipular de acordo com a resolução desejada ora no domínio temporal, ora no de frequências.

Entre os vários tipos existentes de transformadas de *wavelet*, a de *Morlet* (Torrence and Compo, 1998) se apresenta como a ideal no estudo de sinais biológicos, como o EEG, devido à sua semelhança com uma onda ou ondas compostas por senóides moduladas por uma curva Gaussiana.

2.2 Conceitos Computacionais

A extração e o processamento de grandes volumes de informações, complexas ou não, independentes da área de atuação, muitas vezes se tornam impossíveis se feitas de forma quantitativa pelo ser humano. Entretanto, esse problema pode ser solucionado hoje em dia graças aos avanços das técnicas computacionais.

Temos por definição que a Ciência da Computação se destina ao estudo e aplicação de algoritmos. Esses podem ser formulados com base na repetição de operações previamente

observadas no cotidiano ou mesmo baseado em estruturas matemáticas indispensáveis para a resolução de problemas oriundos das mais diversas áreas do conhecimento humano, algumas dessas áreas, tais como a Biologia, produzem grandes volumes de dados todos os dias, sendo muitos desses subaproveitados na aquisição de conhecimento sobre o evento biológico.

O armazenamento e a manipulação de dados podem ser tratados através de vários métodos computacionais já existentes. Entretanto, há na Ciência da Computação uma subárea descrita como “*Knowledge Discovery Database*” (KDD) que pode auxiliar na descoberta de informações relevantes em meio à grande quantidade de dados através da extração semi-automática de conhecimento.

Diferentes metodologias para extração de conhecimento a partir de dados já foram propostas (Fayyad et al., 1996; Weiss and Indurkha, 1996). De forma geral, o processo pode ser descrito em três etapas elementares: (i) pré-processamento, (ii) mineração de dados e (iii) pós-processamento (Baranauskas, 2001).

Na etapa de pré-processamento ocorre a preparação dos dados que são coletados e transformados adequadamente para o início do processo de mineração. Nessa etapa podemos ter a realização de inúmeros processos para limpeza e filtragem dos dados, assim como a manipulação para ressaltar propriedades relevantes e mesmo uma pré-seleção das informações a serem analisadas.

A etapa de mineração tem como objetivo a previsão ou descrição do evento analisado. No caso da predição temos a utilização de valores existentes no conjunto de dados para prever novos valores desconhecidos que podem levar a um evento desejado. Já na descrição temos uma análise dos dados em busca de padrões relevantes que podem auxiliar num melhor entendimento do evento analisado. Para tanto são necessários algoritmos computacionais que utilizam inferência indutiva a partir de um conjunto de dados (ou conjunto de exemplos), e podem ser aplicados em problemas de diversos domínios do conhecimento, entre eles, a

Biociências (Salzberg, 1995;Golub et al., 1999;Tanabe et al., 1999;Mjolsness and DeCoste, 2001;Heckerling et al., 2004).

Por último, na etapa de pós-processamento, ocorre a interpretação dos resultados, que consiste na avaliação dos classificadores extraídos, isto é, o conjunto de regras produzidas para classificação de um exemplo. Nessa etapa ocorre a remoção de padrões irrelevantes ou redundantes e a tradução de padrões úteis em termos inteligíveis pelos especialistas. Também é nessa etapa que ocorre o uso do conhecimento extraído, que consiste na incorporação do conhecimento ao domínio de estudo, seja tomando ações baseadas no conhecimento novo ou simplesmente documentando e relatando para as partes interessadas o que foi obtido, bem como remoção de conflitos potenciais com conhecimento previamente tido como correto.

Um diagrama dos passos necessários para a extração de conhecimento (KDD) pode ser visto na Figura 01.

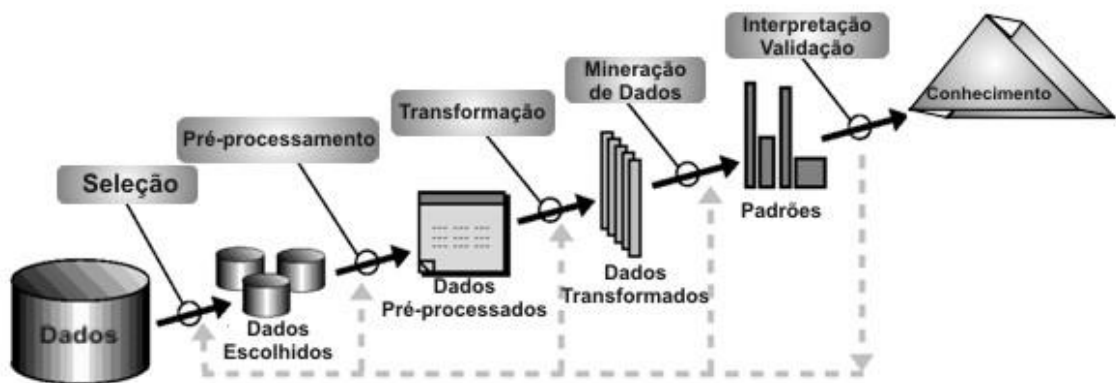


Figura 01 – Etapas do KDD segundo Fayaad (Fayyad et al., 1996)

2.2.1 Mineração de Dados

A segunda etapa descrita para o processo de extração semi-automática de conhecimento (KDD) pode ser obtida através de 05 métodos principais, sendo eles: Classificação, Regressão, Associação, Clusterização e Sumarização.

O método da Classificação ocorre com o uso de algoritmos de indução para realizar a busca por modelos que descrevam ou diferenciem as diversas classes possíveis. Como exemplo, podemos apresentar uma árvore de decisão a ser aplicada para a classificação e conseqüente predição de classe em exemplos futuros com base em valores previamente conhecidos.

Já a Regressão pode ser utilizada, analogamente à Classificação, como um método de predição, entretanto sua aplicação consiste em estimar valores para uma variável contínua desconhecida.

O método da Associação baseia-se na identificação de eventos que possam co-ocorrer em uma mesma transação, ou período. Possibilitando estipular a correlação de atributos na influência de um evento.

O quarto modelo de mineração de dados é conhecido como Clusterização, onde é realizada a partição de um grupo heterogêneo em vários subgrupos mais homogêneos. Essa partição ocorre através da busca pela máxima similaridade entre os conjuntos de exemplos analisados.

Segundo Fayyad (Fayyad et al., 1996), o último modelo de mineração chamado Sumarização, pode ser definido pela busca de uma descrição compacta para um subconjunto de dados. O resultado desse modelo é apresentado em forma de regras de síntese sobre o evento analisado.

De acordo com os modelos citados anteriormente, podemos concluir que diferentes técnicas de mineração de dados podem ser aplicadas para diferentes propósitos, apresentando vantagens e desvantagens de acordo com o contexto onde serão aplicadas. Em domínios onde é necessário validar o conhecimento extraído, como por exemplo na área da saúde, é aconselhável o uso de classificadores simbólicos (Kubat et al., 1972;Dietterich, 1998).

Capítulo 3 – Metodologia e Procedimentos

Esse capítulo abordará os procedimentos utilizados para a obtenção dos dados utilizados na Mineração de Dados. Assim como, o pré-processamento dos dados, a formatação no estilo apropriado e os passos utilizados para a extração semi-automática de conhecimento.

3.1 Obtenção dos Dados

3.1.1 Estimulação

Visando a questão ética do reaproveitamento de dados experimentais, o presente trabalho utilizou registros já existentes provenientes do projeto de Doutorado de Maira Licia Foresti, doutoranda pelo Laboratório de Neurofisiologia e Neuroetologia Experimental – LNNE (processo FAPESP n° 05/51366-9).

A obtenção desses registros foi realizada como parte do doutorado através de um protocolo de abrasamento elétrico rápido da amígdala aprovado pelo Comitê de Ética em Pesquisa Animal da FMRP - USP (número de protocolo 192/2005).

De acordo com o protocolo, ratos Wistar machos foram implantados com eletrodos na amígdala e hipocampo (Foresti et al., 2008). Os animais foram submetidos a estimulações na Amígdala Baso-Lateral (BLA) esquerda através de um sistema de estimulação GRASS S88 (West Warwick, EUA) conectado a um osciloscópio (Hewlett-Packard 130C HP, Palo Alto, EUA) para monitoramento dos sinais enviados aos cérebros dos ratos. As estimulações elétricas seguiram os seguintes parâmetros: estímulo elétrico monopolar com trens de ondas

quadradas bifásicas, corrente constante com frequência de 60 Hz e pulsos com 01 ms de duração.

Inicialmente foi realizado um teste de limiar de Pós-Descarga (PD), onde foi verificada a amperagem mínima para que o animal apresentasse uma crise eletrencefalográfica pós-estímulo. Esse teste consiste no registro do EEG do animal durante 05 minutos e a aplicação do estímulo elétrico por 02 s iniciando em 100 μ A. Caso não haja reação o ciclo é realizado novamente acrescentando se 100 μ A até que se consiga o limiar de PD do animal.

A reação de PD foi definida como a atividade epileptiforme que ocorre imediatamente após o estímulo com pelo menos 01 Hz de frequência e amplitude 02 vezes maior que a linha de base do EEG (Ebert and Löscher, 1995).

Depois de estabelecido o limiar de PD, os animais foram submetidos a um total de 21 estímulos, sendo 10 no primeiro dia, 10 no segundo dia e o último no terceiro dia, todos com duração de 10 segundos e separados por um período de 30 minutos, conforme descrito por (Foresti et al., 2008)

3.1.2 Captura

O sinal eletrofisiológico da BLA esquerda e do hipocampo ipsilateral, foi registrado passando por uma placa eletrônica com 02 canais sem interrupção de sinal (estímulo e terra) e 02 canais ligados a um transistor de efeito de campo (*Field Effect Transistor* – FET; JFET SST201, encapsulamento SOT23), para registros monopolares. Os FETs fazem parte de um sistema *source follower* para redução de artefatos de movimento ou de fontes externas. Esse sistema funciona conectado a duas pilhas tipo AAA de +1,5 V, fazendo assim uma fonte de +1.5V e -1,5V e a resistências SMD de 10 k Ω constituindo um pré-amplificador de ganho 1x.

Esse aparato é montado no cabo que, de um lado se conecta com os eletrodos implantados no encéfalo do animal, e do outro lado se conecta a uma caixa de distribuição de sinais através de um *swivel* de quatro canais. O *swivel* permite a mobilidade do animal na caixa de acrílico onde ocorrem os estímulos, com artefatos de movimento reduzidos ou ausentes. O sinal é conduzido a um sistema de amplificação e filtragem (CyberAmp 320, Axon Instruments, Foster City, EUA) e a um conversor Analógico/Digital (Biopac Systems mod. MP100, Santa Barbara, EUA), acoplado a um computador (Dutra Moraes et al., 2000).

A etapa de captura dos sinais de EEG foi realizada como parte de outro projeto (Dutra Moraes et al., 2000; Foresti et al., 2008) segundo especificações de Dutra Moraes (Dutra Moraes et al., 2000).

3.1.3 Registro

Os animais foram filmados (câmera VHS, Gradiente, GCP-180CR) e as imagens foram enviadas através de uma placa conversora de captura e saída de vídeo (*ATI All-in-Wonder Pro*) para a tela do computador. A imagem da tela do computador, composta pelo traçado EEGráfico e pelos comportamentos do animal foi gravada em fita formato VHS conforme a técnica de vídeo-EEG (Moraes, 1998; Dutra Moraes et al., 2000; Romcy-Pereira and Garcia-Cairasco, 2003; Galvis-Alonso et al., 2004). Todos os registros foram realizados 01 minuto antes e 05 minutos após cada estímulo, durante todo o protocolo de abasamento.

O registro EEG também foi coletado e armazenado digitalmente utilizando o software *Acknowledge* (Biopac Systems, USA), sendo a taxa de amostragem digital para todos os registros de 500 Hz, ou seja, 500 valores de potencial por segundo em cada um dos canais registrados, Amígdala e Hipocampo.

Através do programa *Quick300* (Axon Instruments) foi determinado um ganho total de 1000 vezes, filtro passa-baixa em 10.000 Hz, filtro passa-alta em 0,1 Hz e *notch-filter* de 60 Hz para eliminar a frequência associada à rede elétrica.

3.2 Pré-Processamento

Com base nos dados do grupo experimental disponíveis para esse trabalho, a fase de pré-processamento foi realizada nos registros de 08 animais (n = 08), denominados:

MENR09; MENR10; MENR14; MENR15; MENR25; MENR26; MENR29; MENR31.

Esta fase pode ser dividida em 4 etapas: Análise de Vídeo-EEG; Classificação de Épocas; Extração de Informações e Criação do arquivo Arff, descritas a seguir.

3.2.1 Análise de Vídeo-EEG

Todos os registros foram submetidos à análise de Vídeo-EEG, correlacionando estado comportamental e traçado EEGráfico.

Essa etapa foi realizada primeiramente por Foresti (Dutra Moraes et al., 2000; Foresti et al., 2008) e para efeitos de adequação aos requisitos estabelecidos no presente estudo, também passaram por uma segunda análise onde foi priorizada a análise do comportamento e do traçado EEGráfico. Ambos os resultados foram comparados para uma pré-seleção dos registros a serem utilizados.

Por meio dessa análise foi possível eliminar registros que estivessem contaminados por ruído de artefatos ou mesmo de má captura do sinal. Foi nessa etapa também que estabelecemos o evento de interesse.

Para se evitar a interferência do próprio estímulo elétrico era desejado, como evento de interesse, a entrada do animal no período ictal espontaneamente, o que foi chamado de Pós-Descarga Secundária (PDS). Esse acontecimento foi observado em 100% dos animais, em alguns dos 21 estímulos, diferenciando-se apenas em número de vezes e na ocorrência do evento simultaneamente em ambas as áreas, Amígdala e Hipocampo.

Para o estudo do sinal eletrencefalográfico por meio de análises no domínio espectral foi estipulado que a PDS deveria ter no mínimo 10 segundos de duração e ocorrer no mínimo 05 vezes entre os 21 estímulos para viabilizar o estudo naquele animal específico.

De acordo com as regras estipuladas para a composição da amostra a ser estudada, foram eliminados os registros dos animais MENR10 e MENR25, restando 06 animais (n = 06) para prosseguir com o pré-processamento, conforme exemplifica a Figura 02.

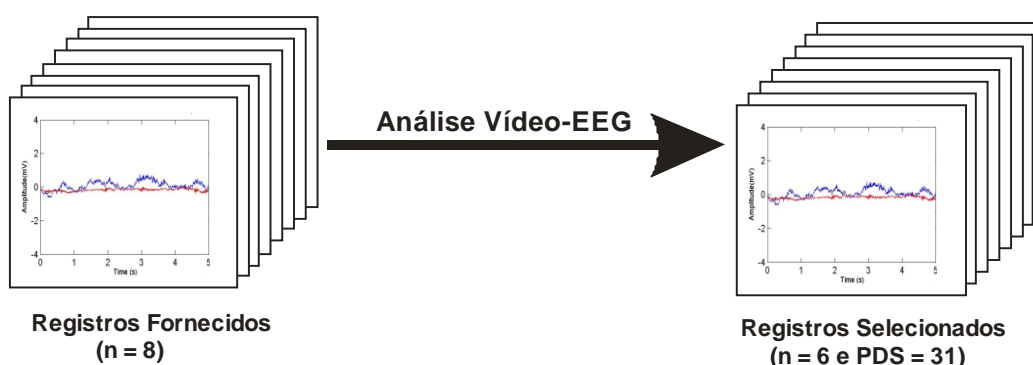


Figura 02 – Seleção dos registros disponíveis para início do estudo.

Do total inicial de 08 registros (animais), 06 foram selecionados como adequados.

A partir dos 06 registros foi possível identificar 31 períodos de Pós-Descarga Secundária.

3.2.2 Classificação de épocas

Após a eliminação dos registros de animais que não obedeciam aos requisitos mínimos estipulados anteriormente, foi necessário desenvolver um algoritmo computacional para separar os períodos referentes às PDS.

Para essa tarefa foi realizada uma nova análise visual dos dados, agora somente pelo traçado EEGráfico, onde foram marcados os tempos de início e de fim das crises secundárias.

O início da PDS foi definido como o tempo, em segundos, onde o traçado apresentou atividade epileptiforme com pelo menos 01 Hz de frequência e amplitude 02 (duas) vezes maior que a linha de base do EEG (Ebert and Löscher, 1995). Já o final da crise secundária foi definido como o tempo onde houve a atenuação do traçado.

Com a marcação dos pontos iniciais e finais das crises, foi possível obter a referência necessária para a separação de épocas, isto é, através de uma rotina desenvolvida para o *software MatLab* (versão 6; *The Math Works*, Natick, EUA) foi realizado o corte do registro somente no período desejado resultando em 02 (dois) sub-registros nomeados de Pré-ictal e Fim-Ictal. Esses correspondem aos 10 segundos anteriores à crise e aos 10 segundos no final da crise, respectivamente.

É importante ressaltar que segundo a nomenclatura sugerida pela literatura, somente períodos anteriores a uma primeira crise são denominados Pré-Ictal. De acordo com essas mesmas definições, o período que sucede a primeira ou demais crises e antecede o próximo período ictal deve ser chamado de inter-ictal. Entretanto, somente para fins de identificação essa nomenclatura foi utilizada para referenciar os dois períodos analisados, Pré-Ictal e Fim-Ictal, ao evento de Pós-Descarga Secundária, e não a todo o histórico da patologia no animal, conforme mostra a Figura 03.

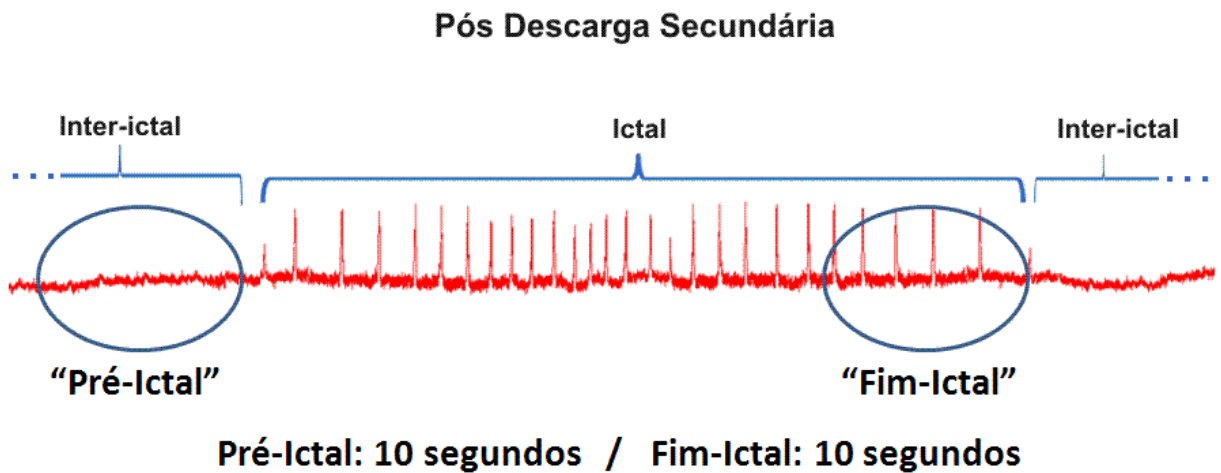


Figura 03 – Esquema ilustrativo do traçado EEGráfico.

Exemplo de registro EEG simulado com traçado epileptiforme e sua devida nomenclatura provida pela literatura e estipulada nesse estudo para relacionar os períodos pré-ictal e fim-ictal ao evento de interesse, PDS.

Ao final da etapa de separação de épocas havia disponível para análise matemática / computacional 124 conjuntos de dados, sendo: 62 sub-registros para Amígdala e 62 sub-registros para Hipocampo, conforme mostrado na Figura 04. Esses subregistros são resultantes do algoritmo desenvolvido (ANEXO 5).

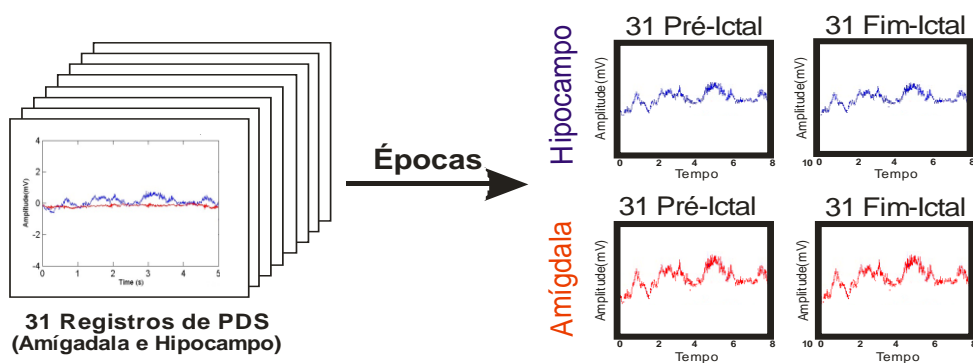


Figura 04 – Resultado do Algoritmo para Separar Épocas.

Número de registros para análise matemática / computacional após a passagem do algoritmo para separação de épocas nos 31 registros de PDS selecionados, dos quais resultaram em 124 registros: 31 Pré-Ictal da Amígdala, 31 Fim-Ictal da Amígdala, 31 Pré-Ictal do Hipocampo e 31 Fim-Ictal do Hipocampo.

3.2.3 Extração de informações

Nessa 3ª etapa do pré-processamento foi desenvolvido um algoritmo a ser utilizado no *software MatLab*, que fosse capaz de extrair o domínio espectral ao longo de um período desejado. Adicionalmente, foi realizado um estudo dos algoritmos já desenvolvidos para análise de frequências, assim como alguns algoritmos necessários para processamento de sinais visando o reaproveitamento de códigos.

O resultado gráfico desse procedimento pode ser visto na Figura 05 onde um sinal é simulado por 10 segundos com frequência de 10 Hz durante os primeiros 05 segundos e 30 Hz durante os 05 últimos segundos. O gráfico é composto pelo sinal amostrado e pela análise espectral por Transformada de Fourier resultante do algoritmo desenvolvido para esse trabalho (ANEXO 6).

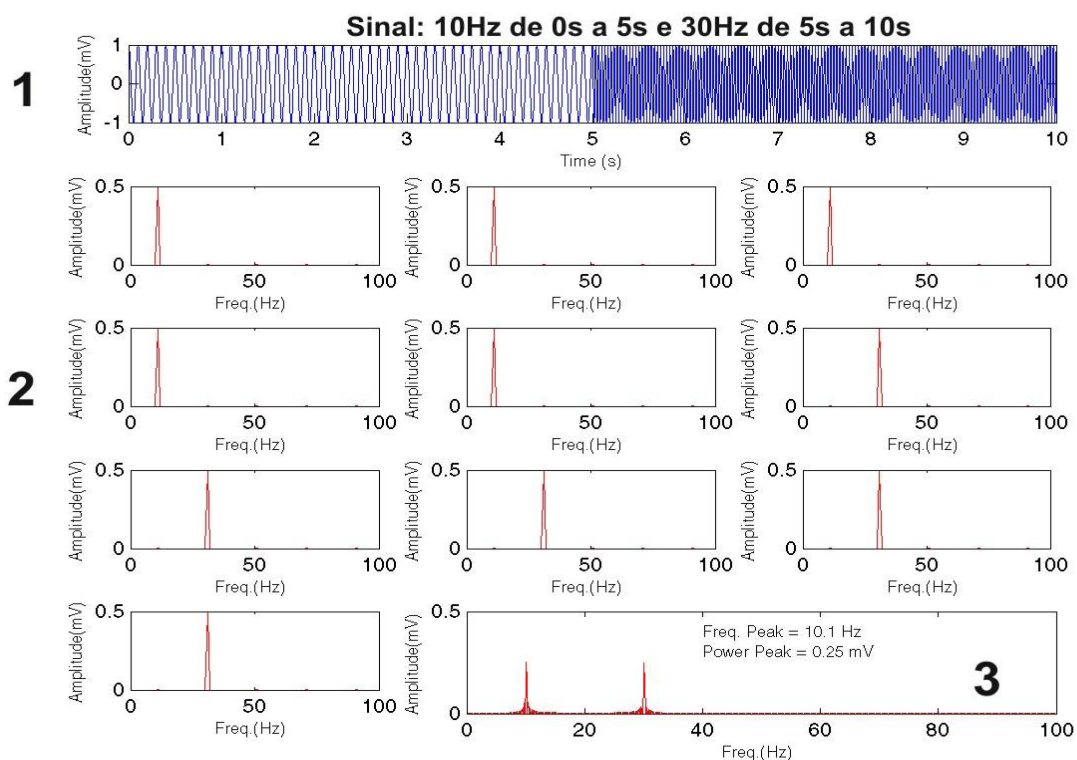


Figura 05 – Diagrama de Sinal Simulado: Análise Espectral por Fourier

1. Gráfico do sinal amostrado ao longo do período. **2.** Conjunto de gráficos no domínio Amplitude vs. Frequência a cada segundo. **3.** Gráfico contendo a análise espectral no período todo relacionando as frequências presentes, porém sem relação de tempo de ocorrência.

A partir da extração do domínio espectrográfico podemos observar a distribuição de frequências e suas respectivas potências (energia) segundo a segundo. Entretanto não temos uma relação temporal com a ocorrência das mesmas.

Para correlacionar o resultado desse algoritmo com o domínio de tempo durante todo o período, foi desenvolvida também uma rotina computacional baseada na Transformada de *Wavelet* para correlação de frequência, potência do sinal e sua localização no tempo, conforme mostra a Figura 06. A partir desse resultado podemos observar o comportamento do sinal EEGráfico no domínio de frequências, o que pode auxiliar num melhor entendimento do evento a ser analisado.

Esse segundo algoritmo implementado (ANEXO 7) foi desenvolvido baseado no código para cálculo da energia de sinais digitais de Ole Jensen (1999) e Dráulio Araujo (2000).

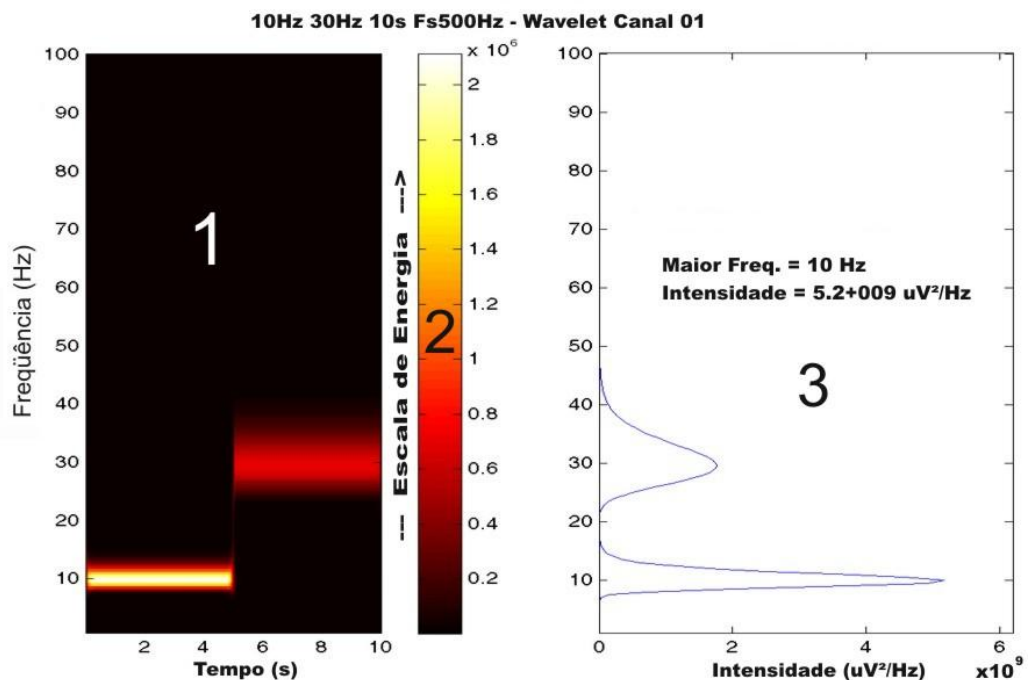


Figura 06 – Diagrama de Sinal Simulado: Análise Espectral por Wavelet

1. Espectrograma resultante da análise por Wavelet do sinal amostrado em 500Hz com oscilação de frequência de 10Hz para 30Hz após 05 seg. 2. Escala de Energia do sinal analisado. 3. Diagrama correlacionando as maiores frequências com suas respectivas intensidades.

Uma vez finalizada a passagem da rotina em cima do sinal digital, é possível obter um vetor de potenciais elétricos com suas respectivas frequências, do qual podemos criar vários esquemas de informações que descrevem o comportamento do sinal de EEG.

Algumas das informações extraídas pelos algoritmos da transformada de fourier e da transformada de wavelet, tornadas relevantes para a Mineração de dados são listadas abaixo:

- 03 e 05 maiores potências e suas respectivas frequências por segundo e no período;
- Médias de potenciais e médias de frequências por segundo e no período;
- Taxa⁴ de crescimento e decrescimento do potencial elétrico do sinal ao longo do período.

3.2.4 Arquivo ARFF

Um terceiro algoritmo foi criado para transformar as informações extraídas no formato adequado para a mineração de dados.

De acordo com a definição fornecida pela ferramenta *Weka* (Witten and Frank, 2005) - *Waikato Environment for Knowledge Analysis* – utilizada para o processo de Mineração de Dados, os conjuntos de dados (também denominados de conjuntos de exemplos), deverão estar no formato apropriado para ser possível identificar o nome de cada atributo, tipo de valor, classe de predição do exemplo e os dados de cada exemplo.

Uma amostra do formato suportado pela ferramenta *Weka* pode ser visto na Figura 07, onde se encontra o nome do conjunto de exemplos na primeira linha identificado pelo parâmetro “@relation”, os nomes de atributos presentes assim como seus respectivos tipos, que podem ser divididos entre numéricos contínuos denominados de “real” e atributos discretos onde temos opções de valores pré-determinadas. O último atributo corresponde às

classes possíveis de serem preditas, contudo, somente é utilizado no caso de treinamento supervisionado. Após as definições dos atributos, temos a seção @data onde especificamos cada conjunto de exemplo a ser aprendido. Esses são construídos com os valores de cada atributo separados por vírgula e finalizados com a classe a ser predita.

```
@relation arquivo_exemplo
@attribute idade real
@attribute altura real
@attribute peso real
@attribute antecedentes {sim, nao}
@attribute classe {sim, nao}

@data
30,175,70,nao,sim
25,170,65,nao,sim
20,183,80,sim,nao
35,172,73,nao,sim
```

Figura 07 – Exemplo de arquivo ARFF para uso na ferramenta Weka.

Com a realização das etapas anteriores foram criados nesse trabalho 07 conjuntos de exemplos baseados nas 03 e 05 maiores potências e frequências extraídas. Essas são medidas em micro Volts (mV) e suas Frequências em Hertz (Hz).

Além desses 07 conjuntos de exemplos, foram também criados outros 02 conjuntos que uniam todas as informações apenas para verificação de algum tipo de relação entre elas durante a mineração de dados. Contudo, esses 02 conjuntos não traziam informações novas para a análise.

Os conjuntos obtidos estão definidos abaixo:

1. **Conjunto 01 – 69 Atributos:** 03 maiores Potências, 03 Frequências correspondentes, por segundo e no período;

⁴ Taxa de Crescimento e Decrescimento é definido como a taxa de aumento (acima de 1,0) ou diminuição (abaixo de 1,0) do potencial (mV) elétrico $\frac{Potencial(t+1)}{Potencial(t)}$ e da frequência $\frac{Frequência(t+1)}{Frequência(t)}$ do sinal, onde t é o tempo de medição.

2. **Conjunto 02 – 113 Atributos:** 05 maiores Potências, 05 Frequências correspondentes, por segundo e no período;
3. **Conjunto 03 – 25 Atributos:** Média das 03 maiores Potências e Média das Frequências correspondentes, por segundo e no período;
4. **Conjunto 04 – 25 Atributos:** Média das 05 maiores Potências e Média das Frequências correspondentes, por segundo e no período;
5. **Conjunto 05 – 69 Atributos:** Média de Potência nas faixas de Frequência sugeridas pela literatura – Delta, Teta, Alfa, Beta, Gama – por segundo e no período;
6. **Conjunto 06 – 12 Atributos:** Taxa de Crescimento / Decrescimento da Potência, por segundo tomando como base média das 03 maiores;
7. **Conjunto 07 – 12 Atributos:** Taxa de Crescimento / Decrescimento da Potência, por segundo tomando como base média das 05 maiores;
8. **Conjunto 08 – 185 Atributos:** União dos dados baseados nas 03 maiores Potências.
9. **Conjunto 09 – 229 Atributos:** União dos dados baseados nas 05 maiores Potências.

Cada conjunto de exemplo também teve como atributo a identificação do animal, a área registrada (Amígdala e Hipocampo) e a classe predita (Pré-Ictal e Fim-Ictal).

É necessário ressaltar que os 05 primeiros conjuntos de exemplos foram criados com base nas informações extraídas pela transformada de fourier, uma vez que não era necessário o conhecimento sobre o tempo de ocorrência das frequências e potências. Já os conjuntos 06 e 07 foram montados com base na extração de informações pela transformada de wavelet, com a qual correlacionamos a ocorrência das frequências e potências com seus respectivos tempos de transição.

3.2.5 Mineração de Dados

A fase de Mineração de Dados (*Data Mining*) foi realizada com o auxílio da ferramenta *Weka*, a qual se destaca entre as outras ferramentas da área por ser de domínio público e permitir uma série de modificações de parâmetros em diversos algoritmos, além de fornecer informações estatísticas sobre o aprendizado.

Para o processo de mineração foi escolhido o método de Classificação. Com esse método podemos obter árvores de decisão que são capazes de classificar novos exemplos e com isso prever o evento de interesse, período Pré-Ictal ou Fim-Ictal, para tanto foram selecionados alguns algoritmos que se destacam na ferramenta *Weka*.

A seleção desses algoritmos foi feita baseando-se em testes preliminares entre diferentes tipos de análise por mineração de dados. Nessas análises foram avaliados algoritmos como os geradores de regras PART e ZeroR, o Naive Bayes, J48 e mesmo abordagens com redes neurais artificiais através da utilização do Multilayer Perceptron, todos disponíveis na plataforma *Weka*.

Todos os testes preliminares mostraram que os melhores desempenhos e a maior simplicidade nos resultados obtidos pertenciam a dois algoritmos, sendo eles, J48 e PART que foram selecionados para se aplicar a técnica computacional de mineração de dados.

O algoritmo J48 é um dos algoritmos de classificação da ferramenta *Weka*. Seu funcionamento consiste na construção de um modelo de Árvore de Decisão com base num conjunto de treinamento.

A construção da árvore é realizada pela escolha de um atributo para servir de referência no teste dos exemplos, isto é, após inserir um conjunto de treinamento, o algoritmo seleciona um atributo para agir como limiar na classificação de novos exemplos, esse é denominado “Nó Raiz” e se torna responsável pelo primeiro teste lógico na classificação de

novos exemplos. Caso seja necessário, novos atributos são escolhidos e tomados como divisores, ou “Nós Intermediários (ou não Folhas)”, cuja função é estabelecer novos testes lógicos e conseqüentes ramificações da árvore. O processo de ramificação ocorre até que não haja exemplos de classes diferentes em uma mesma classificação, a qual é denominada “Nó folha”.

Um exemplo de Árvore de Decisão pode ser visto na Figura 09, onde é apresentado o resultado da classificação do arquivo de dados exemplificado na Figura 08.

A leitura da árvore é feita a partir do atributo escolhido pelo algoritmo como ponto de partida para classificação de exemplos, o que no exemplo foi definido como “Altura”. Através de testes lógicos para cada atributo definimos o caminho a percorrer até a classificação de todos os exemplos. De acordo com a Figura 08 podemos concluir que:

- a) Caso a altura seja menor que 175 cm temos a predição da Classe Sim.
- b) Caso a altura seja igual ou maior que 175 cm e o Peso menor que 75 Kg teremos a Classe Sim.
- c) Caso a altura seja igual ou maior que 175 cm e o Peso igual ou maior que 75 Kg, teremos a Classe Não.

Podemos observar que nesse exemplo houve a exclusão dos atributos idade e antecedentes pois apenas com os atributos Altura e Peso já foi possível cobrir todos os exemplos aprendidos, e com isso descobrir um padrão existente nos mesmos.



Figura 08 – Exemplo de Árvore de Decisão. (Waikato 2004)

Cada forma oval define um teste lógico que classificará o exemplo.

Cada retângulo representa um nó folha com a respectiva classe e o número de exemplos classificados.

O segundo algoritmo escolhido, PART, apresenta-se como uma variante do algoritmo J48 e sua utilização está relacionada à criação de regras de produção e seu refinamento a partir da árvore induzida.

O refinamento das regras é realizado de forma iterativa com base na medida de cobertura, ou seja, quantos exemplos são satisfeitos pela regra construída. A cada refinamento é verificada a cobertura de exemplos, e o processo somente será finalizado quando todos os exemplos forem cobertos.

Capítulo 4 – Resultados

4.1 Análise Computacional

4.1.1 Escolha de parâmetros

A etapa de Mineração de Dados foi focada inicialmente na escolha dos 03 melhores Conjuntos de Exemplos definidos anteriormente (Seção 3.2.4).

Para essa tarefa foi utilizado o algoritmo de indução de árvores de decisão J48 com variação de 02 parâmetros:

- **minNumObj**: Número Mínimo de exemplos em um nó final para esse ser considerado uma folha. Número alto tende à generalização da árvore. Já os baixos à especificação.

- **confidenceFactor**: Fator de confiança para a poda da árvore de decisão. Valores altos minimizam a poda, enquanto os mais baixos a intensificam permitindo apenas nós com significância estatística.

❖ Foram testadas também árvores sem a realização de Poda, denominadas de Árvores Brutas

Os resultados dessas comparações foram analisados através de métricas quantitativas sobre a classificação de exemplos, e assim concluído quais conjuntos apresentaram melhor desempenho. As métricas utilizadas para a análise são:

- Porcentagem de Classificação Correta / Incorreta;
 - Indica quantos exemplos entre os fornecidos foram classificados correta ou incorretamente de acordo com a árvore de decisão.
- Índice *Kappa*: Indica a concordância entre as classes preditas e as verdadeiras.
- Matriz de Confusão e suas métricas, conforme apresentado na Tabela 01:

Tabela 01 – Exemplo de Matriz de Confusão na Classificação de Exemplos.

Classe	Preditos como A	Preditos como B
A	Verdadeiro Positivo Vp	Falso Negativo Fn
B	Falso Positivo Fp	Verdadeiro Negativo Vn

Definições

Vp – Verdadeiro Positivo – Classificar como Classe A exemplos de Classe A.

Fn – Falso Negativo – Classificar como não ser Classe A, ou ser da Classe B exemplos de Classe A.

Fp – Falso Positivo – Classificar como não ser Classe B, ou ser Classe A exemplos de Classe B.

Vn – Verdadeiro Negativo – Classificar como não ser Classe A, ou ser Classe B exemplos de Classe B

- **Sensibilidade:** Probabilidade de se rotular como Não sendo da Classe em

questão e realmente não ser dessa classe $\rightarrow \frac{Vp}{(Vp + Fn)}$

- **Especificidade:** Probabilidade de se rotular como sendo da Classe em

questão e realmente ser dessa classe $\rightarrow \frac{Vn}{(Vn + Fp)}$

4.1.2 Algoritmo J48

O processo de aprendizado foi realizado com base nos conjuntos de exemplos construídos e através da técnica de validação cruzada (*cross-validation*) (*folds* = 10).

Essa técnica permite dividir o conjunto de exemplos em tantas vezes quanto o número de *folde* estabelecido. Cada *fold* é utilizado no teste de predição pela árvore construída pelo restante dos exemplos. Com isso temos então nesse estudo a divisão de cada conjunto de exemplos em 10 partes que serão testadas por 10 diferentes árvores de classificação.

Essa repetição nos processos de treinamento e teste aumenta as estimativas de predição.

O resultado da comparação dos conjuntos de exemplos através do algoritmo J48 variando o *minNumObj* em 01, 02 e 03 exemplos, sem poda (Árvore Bruta) e com poda (Árvore Podada) utilizando 0.10, 0.25 e 0.50 como Fator de Confiança, pode ser visto abaixo através das Tabelas 02, 03, 04 e 05.

Tabela 02 – Representação das classificações corretas do J48, em porcentagem.

Tabela comparativa da Classificação Correta em Porcentagem usando o algoritmo J48 com *minNumObj* = 01, sem poda (Bruta) e com poda (Podada) usando o fator de confiança em 0.10, 0.25 e 0.50.

Classificação Correta dos Exemplos em Porcentagem (%) – minNumObj = 01				
Conjunto	Classificação (%) com desvio padrão			
	Bruta	Podada (0.10)	Podada (0.25)	Podada (0.50)
Conjunto 01	85,06 ± 11,73	85,85 ± 11,95	85,94 ± 11,72	85,77 ± 11,93
Conjunto 02	84,56 ± 09,81	85,83 ± 10,09	85,53 ± 10,22	84,56 ± 09,81
Conjunto 03	79,46 ± 10,02	81,62 ± 09,80	80,58 ± 09,61	79,54 ± 09,75
Conjunto 04	84,31 ± 08,90	84,88 ± 09,68	84,39 ± 09,36	84,40 ± 08,91
Conjunto 05	83,04 ± 09,52	85,95 ± 10,79	83,13 ± 10,19	83,04 ± 09,61
Conjunto 06	94,85 ± 09,20	94,67 ± 07,84	94,90 ± 07,49	95,38 ± 07,55
Conjunto 07	89,54 ± 09,64	89,54 ± 09,26	90,35 ± 09,43	90,19 ± 09,65
Conjunto 08	86,60 ± 10,24	86,67 ± 10,40	86,51 ± 10,32	86,60 ± 10,26
Conjunto 09	83,31 ± 09,68	83,65 ± 09,49	83,15 ± 09,57	83,47 ± 09,65

Tabela 03 – Representação das classificações corretas do J48, em porcentagem.

Tabela comparativa da Classificação Correta em Porcentagem usando o algoritmo J48 com minNumObj = 02, sem poda (Bruta) e com poda (Podada) usando o fator de confiança em 0.10, 0.25 e 0.50.

Classificação Correta dos Exemplos em Porcentagem (%) – minNumObj = 02				
Conjunto	Classificação (%) com desvio padrão			
	Bruta	Podada (0.10)	Podada (0.25)	Podada (0.50)
Conjunto 01	85,97 ± 11,21	86,46 ± 11,36	86,54 ± 10,91	86,21 ± 11,31
Conjunto 02	84,95 ± 08,93	85,89 ± 09,10	85,66 ± 09,29	85,11 ± 08,99
Conjunto 03	81,37 ± 10,68	82,23 ± 10,53	82,31 ± 10,09	81,36 ± 10,84
Conjunto 04	84,83 ± 09,82	85,30 ± 09,84	84,74 ± 09,69	84,82 ± 09,51
Conjunto 05	81,52 ± 09,35	84,51 ± 10,69	82,74 ± 10,26	81,60 ± 09,76
Conjunto 06	87,17 ± 10,45	86,90 ± 09,78	87,31 ± 09,56	87,63 ± 09,65
Conjunto 07	87,09 ± 09,16	86,68 ± 09,14	87,40 ± 09,23	87,41 ± 09,33
Conjunto 08	86,62 ± 09,87	86,46 ± 10,07	86,46 ± 10,07	86,54 ± 10,00
Conjunto 09	83,33 ± 09,81	83,33 ± 09,60	83,25 ± 09,56	83,17 ± 09,73

Tabela 04 – Representação das classificações corretas do J48, em porcentagem.

Tabela comparativa da Classificação Correta em Porcentagem usando o algoritmo J48 com minNumObj = 03, sem poda (Bruta) e com poda (Podada) usando o fator de confiança em 0.10, 0.25 e 0.50.

Classificação Correta dos Exemplos em Porcentagem (%) – minNumObj = 03				
Conjunto	Classificação (%) com desvio padrão			
	Bruta	Podada (0.10)	Podada (0.25)	Podada (0.50)
Conjunto 01	85,88 ± 11,62	86,28 ± 11,35	86,36 ± 11,53	86,03 ± 11,78
Conjunto 02	85,18 ± 07,98	85,90 ± 08,20	85,73 ± 08,08	85,18 ± 07,98
Conjunto 03	80,15 ± 10,16	81,91 ± 09,97	81,60 ± 09,96	80,38 ± 10,21
Conjunto 04	85,13 ± 09,97	85,51 ± 10,02	85,60 ± 10,35	85,28 ± 09,97
Conjunto 05	82,73 ± 09,89	85,15 ± 10,18	84,76 ± 10,65	83,04 ± 10,09
Conjunto 06	84,28 ± 11,50	83,77 ± 11,06	84,65 ± 11,04	84,82 ± 11,08
Conjunto 07	85,97 ± 09,68	85,63 ± 0,63	86,04 ± 09,66	86,20 ± 09,61
Conjunto 08	86,72 ± 10,04	86,65 ± 09,95	86,64 ± 10,10	86,80 ± 10,05
Conjunto 09	84,00 ± 10,28	83,58 ± 10,35	83,83 ± 10,30	84,08 ± 10,11

Tabela 05 – Melhor Desempenho de acordo com o J48 – minNumObj.

Disposição em ordem crescente dos Conjuntos que apresentaram melhor desempenho na classificação do J48, utilizando minNumObj de 01, 02 e 03 exemplos por folha, sem poda (B.) e com poda (P.) em 0.10, 0.25 e 0.50 no Fator de Confiança.

Conjunto de Melhor Desempenho em Classificação Correta					
J48 – Com Poda e fator de Confiança de 0.10					
minNumObj = 01 / 02 / 03					
Conjunto	B. (%)	Conjunto	B. (%)	Conjunto	B. (%)
Conjunto 06	94,85 ± 09,20	Conjunto 06	87,17 ± 10,45	Conjunto 08	86,72 ± 10,04
Conjunto 07	89,54 ± 09,64	Conjunto 07	87,09 ± 09,16	Conjunto 07	85,97 ± 09,68
Conjunto 08	86,60 ± 10,24	Conjunto 08	86,62 ± 09,87	Conjunto 06	84,28 ± 11,50
Conjunto	P. (%)	Conjunto	P. (%)	Conjunto	P. (%)
Conjunto 06	94,67 ± 07,84	Conjunto 06	86,90 ± 09,78	Conjunto 08	86,65 ± 09,95
Conjunto 07	89,54 ± 09,26	Conjunto 07	86,68 ± 09,14	Conjunto 07	85,63 ± 0,63
Conjunto 08	86,67 ± 10,40	Conjunto 08	86,46 ± 10,07	Conjunto 06	83,77 ± 11,06

A análise do Índice *Kappa* indica se uma classificação ou predição concorda com a classe determinada inicialmente. O valor de *Kappa* varia em [-1,1], indicando ausência de concordância em [-1,0] e presença de concordância em [0,1].

Tabela 06 – Representação dos Índices *Kappa* na classificação do J48.

Tabela comparativa dos Índices *Kappa* para as Classificações realizadas com o algoritmo J48 com 01, 02 e 03 minNumObj, sem poda (Bruta) e com poda (Podada) usando o fator de confiança em 0.10, 0.25 e 0.50.

Conjunto	Kappa [-1,1] – 01		Classificação [-1,1] - 02		Classificação [-1,1] - 03	
	Bruta	Podada (0.10)	Bruta	Podada (0.10)	Bruta	Podada (0.10)
Conjunto 01	0,70 ± 0,23	0,72 ± 0,24	0,72 ± 0,22	0,73 ± 0,23	0,72 ± 0,23	0,73 ± 0,23
Conjunto 02	0,69 ± 0,20	0,72 ± 0,20	0,70 ± 0,18	0,72 ± 0,18	0,70 ± 0,16	0,72 ± 0,16
Conjunto 03	0,59 ± 0,20	0,63 ± 0,20	0,63 ± 0,21	0,64 ± 0,21	0,60 ± 0,20	0,64 ± 0,20
Conjunto 04	0,69 ± 0,18	0,70 ± 0,19	0,70 ± 0,20	0,71 ± 0,20	0,70 ± 0,20	0,71 ± 0,20
Conjunto 05	0,66 ± 0,19	0,72 ± 0,22	0,63 ± 0,19	0,69 ± 0,21	0,65 ± 0,20	0,70 ± 0,20
Conjunto 06	0,90 ± 0,18	0,89 ± 0,16	0,74 ± 0,21	0,74 ± 0,20	0,69 ± 0,23	0,67 ± 0,22
Conjunto 07	0,79 ± 0,19	0,79 ± 0,18	0,74 ± 0,18	0,73 ± 0,18	0,72 ± 0,19	0,71 ± 0,19
Conjunto 08	0,73 ± 0,20	0,73 ± 0,21	0,73 ± 0,20	0,73 ± 0,20	0,73 ± 0,20	0,73 ± 0,20
Conjunto 09	0,67 ± 0,19	0,67 ± 0,19	0,67 ± 0,19	0,67 ± 0,19	0,68 ± 0,20	0,67 ± 0,20

Tabela 07 – Melhores Índices *Kappa* na Classificação do J48
 Representação dos Conjuntos com melhor Índice *Kappa* após passagem do algoritmo J48 com variações de minNumObj e Fator de Confiança na poda das árvores de decisão.

Conjunto de Melhor Desempenho com Relação ao Índice Kappa					
Kappa [-1,1] - minNumObj 1		Kappa [-1,1] - minNumObj 2		Kappa [-1,1] - minNumObj 3	
Conjunto	B. [-1,1]	Conjunto	B. [-1,1]	Conjunto	B. [-1,1]
Conjunto 06	0,90 ± 0,18	Conjunto 06	0,74 ± 0,21	Conjunto 08	0,73 ± 0,20
Conjunto 07	0,79 ± 0,19	Conjunto 07	0,74 ± 0,18	Conjunto 07	0,72 ± 0,19
Conjunto 08	0,73 ± 0,20	Conjunto 08	0,73 ± 0,20	Conjunto 01	0,72 ± 0,23
Conjunto	P. [-1,1]	Conjunto	P. [-1,1]	Conjunto	P. [-1,1]
Conjunto 06	0,89 ± 0,16	Conjunto 06	0,74 ± 0,20	Conjunto 08	0,73 ± 0,20
Conjunto 07	0,79 ± 0,18	Conjunto 07	0,73 ± 0,18	Conjunto 07	0,71 ± 0,19
Conjunto 08	0,73 ± 0,21	Conjunto 08	0,73 ± 0,20	Conjunto 01	0,73 ± 0,23

Após a comparação dos conjuntos de exemplos através do algoritmo J48 com número mínimo por folha de 01, 02 e 03 exemplos, sem poda e com poda da árvore de decisão em 0.10, 0.25 (padrão) e 0.50 no fator de confiança, constatamos que apenas no Conjunto 06 encontramos diferença significativa entre as árvores resultantes nas variações do número mínimo de exemplos por folha e nas de fator de confiança. A falta de significância nos demais exemplos foi constatada através do teste estatístico de *t-Student* com $\alpha = 5\%$ (0,05).

Descartando a diferença significativa pode se concluir também que os melhores conjuntos nas classificações em geral foram os de número 06, 07 e 08.

Esses eram definidos como:

- 06** – “Taxa de crescimento e decrescimento da Potência com base nas 03 maiores potências por segundo.”
- 07** – “Taxa de crescimento e decrescimento da Potência com base nas 05 maiores potências por segundo.”
- 08** – “Todos os dados extraídos baseados nas 03 maiores Potências.”

De acordo com essa classificação, foi extraída a Matriz de Confusão para análise das medidas de Sensibilidade e Especificidade na predição. Essa pode ser vista abaixo na Tabela 08.

Tabela 08 – Representação da Sensibilidade e da Especificidade do J48

Métricas da Matriz de Confusão - J48		
Conjunto	Sensibilidade	Especificidade
Conjunto 01	0,85 ± 0,16	0,88 ± 0,13
Conjunto 02	0,86 ± 0,15	0,88 ± 0,12
Conjunto 03	0,81 ± 0,15	0,83 ± 0,12
Conjunto 04	0,82 ± 0,16	0,88 ± 0,12
Conjunto 05	0,86 ± 0,16	0,88 ± 0,13
Conjunto 06	0,93 ± 0,12	0,97 ± 0,08
Conjunto 07	0,90 ± 0,14	0,90 ± 0,11
Conjunto 08	0,87 ± 0,14	0,87 ± 0,12
Conjunto 09	0,85 ± 0,14	0,84 ± 0,11

De acordo com essas medidas podemos concluir que, entre os conjuntos de exemplos extraídos no estudo, os esquemas de análise 06 e 07 desenvolvidos a partir do taxa de crescimento e decrescimento da potência do sinal foram os que melhor apresentaram desempenho, esse pode ser interpretado como:

- ❖ 93% dos Exemplos de Período Pré-Ictal realmente foram preditos como sendo de tal período.
- ❖ 97% dos Exemplos de Período Fim-Ictal realmente foram preditos como sendo de tal período.

O Conjunto 08 embora apresente o 3º lugar em melhor desempenho, não deve ser considerado, pois apresenta a união de todos os esquemas baseados nas 03 maiores potências por segundo e no período. Isto é, sua classificação com relação ao desempenho pode ser influenciada pelo conjunto 06.

4.1.3 Algoritmo PART

A partir desse ponto foi possível focar os novos testes nos conjuntos de exemplos de melhor desempenho, conjuntos 06 e 07, ambos compostos pela taxa de crescimento e decrescimento da potência e da frequência do sinal EEGráfico. O uso do algoritmo PART foi então considerado para extrair regras de produção que viabilizassem o entendimento das árvores induzidas (ANEXO 1, 2).

Conforme descrito anteriormente, o algoritmo PART faz uso da árvore de decisão construída pelo mecanismo do J48 para então definir regras de produção que possam levar ao evento de interesse, ou classe de predição.

Para efetuar a extração das regras foram testados os mesmos parâmetros da etapa anterior, porém somente utilizando os conjuntos de exemplos 06 e 07. Também foi considerado o número de regras geradas como observado nas Tabelas 9 e 10.

Tabela 09 – Desempenho do PART com os Conjuntos 06 e 07 – Classificação Correta.

Valores da medida “Classificação correta” obtidos pelo algoritmo PART analisando os Conjuntos 06 e 07 com variação de minNumObj, sem poda (Bruta) e com poda(0.1, 0.25, 0.5)

Métricas da Classificação Correta PART				
minNumObj = 1				
Conjunto	Bruta	0.10	0.25	0.50
Conjuntos 06	90.71 ± 12,20	91.80 ± 10,05	92.01 ± 10,84	92.01 ± 10,84
Conjuntos 07	86.38 ± 11,73	86.50 ± 12,11	86.75 ± 11,56	86.75 ± 11,56
minNumObj = 2				
Conjunto	Bruta	0.10	0.25	0.50
Conjuntos 06	86.87 ± 12,24	88.06 ± 09,75	87.72 ± 10,46	87.72 ± 10,46
Conjuntos 07	84.24 ± 11,69	84.83 ± 11,20	84.50 ± 11,30	84.50 ± 1130
minNumObj = 3				
Conjunto	Bruta	0.10	0.25	0.50
Conjuntos 06	86.50 ± 10,75	87.72 ± 09,17	87.31 ± 09,89	87.31 ± 09,89
Conjuntos 07	85.06 ± 10,59	83.67 ± 10,99	84.40 ± 11,47	84.40 ± 11,47

Tabela 10 – Desempenho do PART com os Conjuntos 06 e 07 – F-Measure.

Valores da medida “F-Measure” obtidos pelo algoritmo PART analisando os Conjuntos 06 e 07 com variações de minNumObj (1,2,3), sem poda (B) e com fator de poda(0.10, 0.25, 0.5)

*F-Measure → Média Harmônica [0,1] entre a Sensibilidade e Especificidade.

Métricas da Matriz de Confusão – PART				
minNumObj = 1				
Conjunto	Bruta	0.10	0.25	0.50
Conjuntos 06	0.90 ± 0,14	0.91 ± 0,12	0.91 ± 0,13	0.91 ± 0,13
Conjuntos 07	0.86 ± 0,13	0.86 ± 0,13	0.86 ± 0,12	0.86 ± 0,12
minNumObj = 2				
Conjunto	Bruta	0.10	0.25	0.50
Conjuntos 06	0.87 ± 0,13	0.88 ± 0,11	0.88 ± 0,11	0.88 ± 0,11
Conjuntos 07	0.84 ± 0,12	0.85 ± 0,12	0.84 ± 0,12	0.84 ± 0,12
minNumObj = 3				
Conjunto	Bruta	0.10	0.25	0.50
Conjuntos 06	0.87 ± 0,11	0.87 ± 0,10	0.87 ± 0,12	0.87 ± 0,12
Conjuntos 07	0.85 ± 0,12	0.84 ± 0,11	0.85 ± 0,11	0.85 ± 0,11

Em todos os testes executados com o uso do algoritmo PART foram definidos conjuntos de 07 regras de produção, excluindo assim essa medida como diferenciador de desempenho.

4.1.4 Considerações Finais

O método existente na ferramenta *Weka* para seleção supervisionada automática de atributos com base em probabilidade foi considerado durante as análises, contudo não mostrou melhoria no desempenho dos classificadores. Em alguns casos esse mesmo método de filtragem causou uma queda nas classificações corretas indicando exclusão de atributos importantes para o processo de mineração.

Entretanto, como pode ser constatado através dos resultados obtidos, embora houvesse uma pequena diferença entre classificações brutas e podadas, constatamos que a poda, ou seja, a seleção de atributos realizada na construção de árvores de decisão realmente aumentou o desempenho da mineração de dados.

Com base nos resultados das medidas analisadas de ambos os classificadores, J48 e PART, e considerando o número fixo de regras geradas em 07, constatamos que os parâmetros que levam ao melhor desempenho na mineração de dados desse tipo de evento são:

- ✓ Classificação pelo Algoritmo J48
- ✓ Conjunto de exemplos composto por taxa de Crescimento e Decrescimento de Potência.
- ✓ $\text{minNumObj} = 1$ objeto mínimo por folha de classificação
- ✓ Árvore resultante após Poda com fator de confiança em 0.10
- ✓ Interpretação por regras de produção definidas pelo PART.

Os resultados das regras induzidas utilizando os Datasets 06 e 07 podem ser vistos nos Anexos 03 e 04 e interpretadas como:

“Pot_Proporcao_3_2 \leq 150 AND Pot_Proporcao_9_8 $>$ 120 Fim-Ictal (20.0)”

Se ocorrer a diminuição da potência em até 50% entre o segundo 02 e o 03 do período analisado E houver aumento de 20% da Potência entre os segundos 08 e 09 conseguimos prever que a crise está terminando.

Cobertura da regra de 20 casos em 124.

4.1.5 Validação da Metodologia

Após a análise de todos os resultados obtidos durante esse trabalho, foram fixadas então as regras e os melhores parâmetros que devem ser obedecidos para se conseguir a predição do evento neurofisiológico em questão.

Nessa etapa de fixação dos requisitos, foi verificada também a diferenciação entre o período ictal e o interictal, isto é, o período entre as crises, porém fora da janela de 10 segundos que referenciamos nesse estudo como pré-ictal.

A diferenciação foi verificada inicialmente com a busca também de assinaturas próprias no período entre crises, contudo, não houve resultado sobre padrões significativos.

Uma segunda tentativa de diferenciação entre os períodos foi realizada com a busca pelas assinaturas obtidas nesse trabalho em períodos que foram, através da análise visual do EEG, identificados como interictais. Essa busca constatou que não há ocorrência das assinaturas fora dos períodos ictal e pré-ictal.

O próximo passo do estudo seria então a validação da metodologia de predição utilizando outros modelos de epilepsia experimental. Para isso foram eleitos 02 modelos realizados também no LNNE.

O primeiro modelo avaliado foi realizado durante o Doutorado de Cristiane Queixa Tilelli (Tilelli et al., 2005) e analisado posteriormente durante o Mestrado de Artur Fernandes (Fernandes, 2008). Esse modelo consistia em dados produzidos pelo modelo de estimulação elétrica da amígdala para indução de *Status Epilepticus (SE)*.

Sua escolha se deu principalmente pela semelhança com o abrasamento elétrico realizado por Maira Lícia Foresti. Entretanto, na primeira fase da metodologia de predição, ou seja, durante o pré-processamento do sinal de EEG,

ocorreram certas dificuldades quanto à diferenciação entre o período basal e o período ictal.

Em todos os registros fornecidos era possível identificar nitidamente o traçado EEGráfico epileptiforme, porém os períodos inter-ictais que seriam utilizados para identificação do início da crise, quando identificados visualmente, se apresentavam com menos de 5 segundos, parâmetro esse estipulado em 10 segundos no mínimo.

Por não apresentar esse parâmetro mínimo não foi possível utilizar os dados produzidos por Cristiane Tilelli e Artur Fernandes para a validação da metodologia de predição.

O segundo modelo avaliado foi realizado pelo Técnico Flávio Del Vecchio como colaborador ao projeto de I.C. de Erika Horta. Esse projeto consiste no estudo da indução ao *Status Epilepticus* por administração intra-hipocampal de pilocarpina.

Os dados fornecidos desse segundo modelo aparentavam ser ideais para a validação da metodologia de predição devido à presença de crises recorrentes espontâneas durante a análise. A ocorrência dessas crises espontâneas possibilitaria uma analogia ao acontecimento das PDSs, crises definidas como evento de interesse na predição.

O pré-processamento dos registros foi realizado conforme estipulado na metodologia, contudo foi possível identificarmos visual e computacionalmente padrões diferentes de basal nos registros fornecidos.

A diferença se encontrava entre o basal controle, isto é, sem nenhum estímulo e o basal do animal após a injeção do veículo da droga. A partir desse ponto seria necessário desenvolver um método de filtragem para normalização do sinal de EEG, ou então uma identificação de períodos basais diferentes.

Qualquer uma das abordagens, seja a identificação do “ruído” causado pelo veículo para realizar a filtragem, seja o reconhecimento de padrões para outros eventos que não a entrada e saída de uma crise, já justificariam por si só outros dois estudos voltados para as áreas de biomédicas e computacional.

De acordo com os dados disponíveis no LNNE e dos critérios estipulados para se aplicar a metodologia de predição, não foi possível realizar a validação em outros modelos de epilepsia experimental. Porém, podemos concluir que uma vez desenvolvida técnica para filtragem do sinal desse segundo modelo, ainda podemos identificar padrões na ocorrência dos traçados epileptiformes. Ainda, como mencionado anteriormente, a validação em modelos experimentais justificaria a busca pela sua implementação na clínica neurológica.

Capítulo 5 – Conclusão

Com base no estudo realizado durante a execução desse projeto vemos que são possíveis diferentes abordagens para análise de sinais eletrencefalográficos sejam eles, visuais, por métodos matemáticos ou mesmo por processamento computacional.

Muitos desses métodos auxiliam na descoberta de propriedades intrínsecas que podem melhorar o entendimento de uma patologia como a Epilepsia, porém a quantidade de dados é tão grande que acabamos por subaproveitar informações provenientes de diversos experimentos.

Vimos também que através de algoritmos desenvolvidos para o processamento de sinais e o uso de algoritmos já existentes para a extração semi-automática de informações conseguimos entender melhor o sinal de EEG e conseqüentemente identificar eventos neurofisiológicos como os disparos anômalos de crises epilépticas.

Com base nesse tipo de identificação é que traçamos o objetivo desse estudo na busca por assinaturas ou padrões EEGráficos que revelem o início e o fim de um período ictal. Concluímos com isso que possuindo o conjunto certo de atributos é possível prever com 93% a 97% de confiança a entrada e saída desse período.

Para finalizar esse trabalho, foi sugerida a validação desses métodos em outros modelos de epilepsia experimental. Entretanto, os registros dos mesmos se apresentavam como inadequados segundo os critérios estabelecidos no início desse estudo, levantando assim as limitações na predição do evento de interesse.

A partir dessas conclusões enfatizamos a necessidade de outras análises seja em processamento de sinais, mineração de dados ou mesmo de outras áreas do conhecimento humano na busca por informações que sejam relevantes e auxiliem no entendimento de problemas biológicos.

Bibliografia

Baranauskas, J. A. 2001. Extração Automática de Conhecimento Utilizando Múltiplos Indutores. ICMC.

Ref Type: Thesis/Dissertation

Ben Ari, Y., 1985. Limbic Seizure and Brain-Damage Produced by Kainic Acid - Mechanisms and Relevance to Human Temporal-Lobe Epilepsy. *Neuroscience* 14, 375-403.

Berger, H., 1929. Über das Elektrenkephalogramm des Menschen (On the human electroencephalogram). *Archiv of Psychiatrie u Nervenkrankheiten* 87 527-570.

Bragin, A., Azizyan, A., Almajano, J., Wilson, C.L., Engel, J., Jr., 2005a. Analysis of chronic seizure onsets after intrahippocampal kainic acid injection in freely moving rats. *Epilepsia* 46, 1592-1598.

Bragin, A., Wilson, C.L., Fields, T., Fried, I., Engel, J., Jr., 2005b. Analysis of seizure onset on the basis of wideband EEG recordings. *Epilepsia* 46 Suppl 5, 59-63.

Dietterich, T.G., 1998. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Comput* 10, 1895-1923.

Dutra Moraes, M.F., Galvis-Alonso, O.Y., Garcia-Cairasco, N., 2000. Audiogenic kindling in the Wistar rat: a potential model for recruitment of limbic structures. *Epilepsy Res* 39, 251-259.

Ebert, U., Löscher, W., 1995. Differences in mossy fibre sprouting during conventional and rapid amygdala kindling of the rat. *Neurosci Lett* 190, 199-202.

Engel, J., Jr., Driver, M.V., Falconer, M.A., 1975. Electrophysiological correlates of pathology and surgical results in temporal lobe epilepsy. *Brain* 98, 129-156.

Fayyad, U., Piatetsky-Shapi, G., Smyth, P., 1996. The KDD Process for Extracting Useful Knowledge from Volumes of Data. *Communications Of The ACM* 39, 27-34.

Fernandes, A. 2008. Expressão aumentada de NPY após Status Epilepticus induzido por estimulação elétrica do complexo amigdalóide em ratos: provável participação em um sistema anticonvulsivo endógeno.

Ref Type: Thesis/Dissertation

Foresti, M.L., Arisi, G.M., Fernandes, A., Tilelli, C.Q., Garcia-Cairasco, N., 2008. Chelatable zinc modulates excitability and seizure duration in the amygdala rapid kindling model. *Epilepsy Research* 79, 166-172.

Fourier, J.B.J., 1822. The Analytical Theory of Heat. *Mémoires de l'Académie Royale des Sciences de l'Institut de France* VII.

Galvis-Alonso, O.Y., Cortes De Oliveira, J.A., Garcia-Cairasco, N., 2004. Limbic epileptogenicity, cell loss and axonal reorganization induced by audiogenic and amygdala kindling in wistar audiogenic rats (WAR strain). *Neuroscience* 125, 787-802.

Garcia-Cairasco, N., Doretto, M.C., Ramalho, M.J., ntunes-Rodrigues, J., Nonaka, K.O., 1996. Audiogenic and audiogenic-like seizures: locus of induction and seizure severity determine postictal prolactin patterns. *Pharmacol Biochem Behav* 53, 503-510.

Goddard, G.V., 1967. Development of epileptic seizures through brain stimulation at low intensity. *Nature* 214, 1020-1021.

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., Lander, E.S., 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286, 531-537.

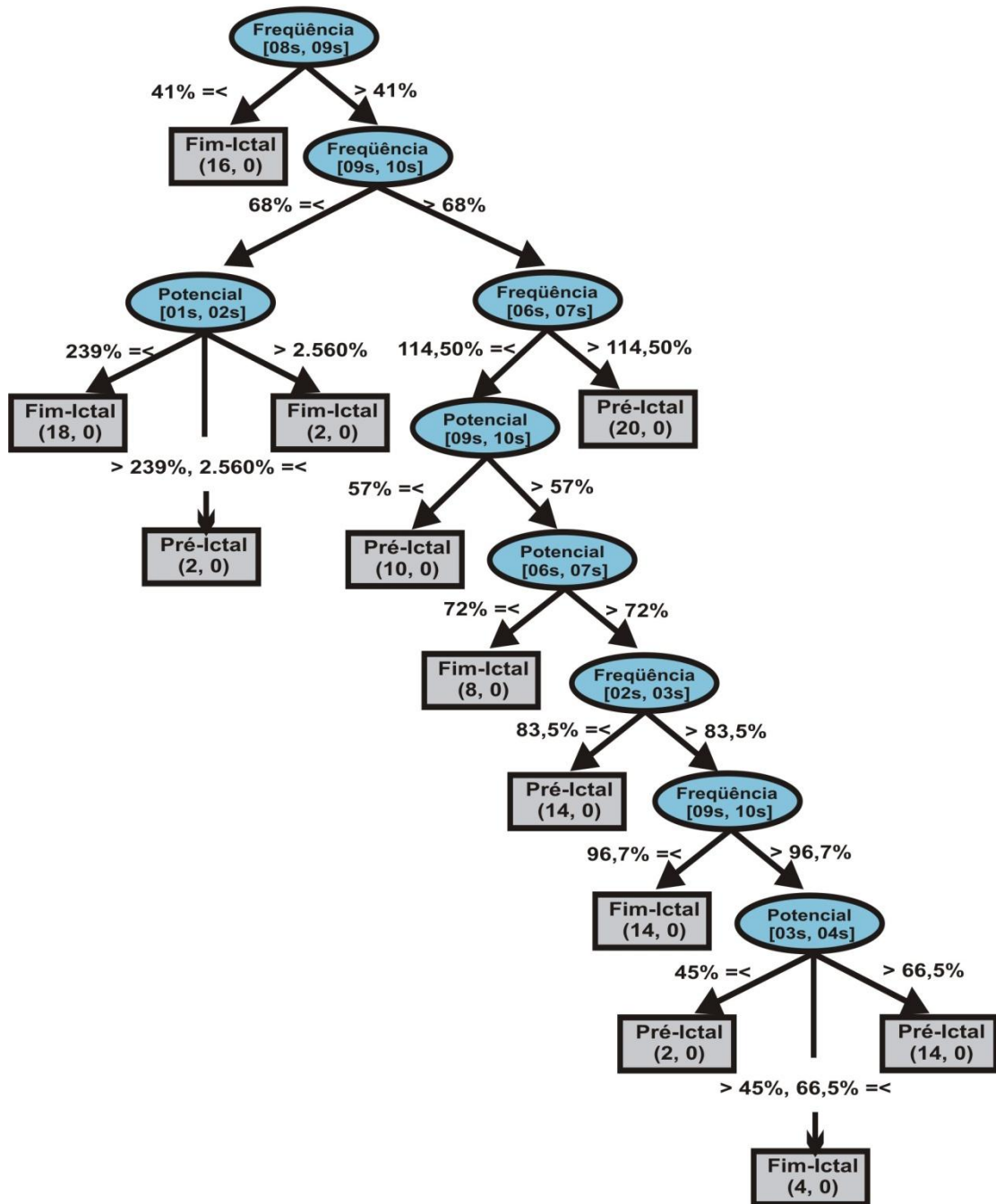
- Grinsted,A., Moore,J.C., Jevrejeva,S., 2004. Application of the cross wavelet transform and wavelet coherence to geophysical time series. *Nonlinear Processes in Geophysics* 11, 561-566.
- Heckerling,P.S., Gerber,B.S., Tape,T.G., Wigton,R.S., 2004. Use of genetic algorithms for neural networks to predict community-acquired pneumonia. *Artificial Intelligence in Medicine* 30, 71-84.
- Houdart, R., Dondey,M., Foncin,J.F., Dreyfus-Brisac,C., 1952. [Traumatic epilepsy with positive spikes; electroencephalogram and corticogram.] . *Rev Neurol (Paris)* 86, 159-160.
- Jasper, H. H. and Droogleever-Fortuyn, J. 1949. Experimental studies on the functional anatomy of petit-mal epilepsy. *Res.Publ.Ass.Nerv.Ment.Dis.* 26, 272-298.
- Kasteleijn-Nolst Trenite,D.G., 1989. Photosensitivity in epilepsy. Electrophysiological and clinical correlates. *Acta Neurol Scand Suppl* 125, 3-149.
- Kiyimik,M.K., Guler,I., Dizibuyuk,A., Akin,M., 2005. Comparison of STFT and wavelet transform methods in determining epileptic seizure activity in EEG signals for real-time application. *Comput Biol Med* 35, 603-616.
- Kubat,K., Stika,L., Nekvinda,V., 1972. Automated processing of prescriptions and methodological guidance. *Cesk Zdrav* 20, 261-268.
- McNamara,J.O., 1984. Kindling: an animal model of complex partial epilepsy. *Ann Neurol* 16, S72-S76.
- Meeren,H.K., Pijn,J.P., van Luijtelaaar,E.L., Coenen,A.M., Lopes da Silva,F.H., 2002. Cortical focus drives widespread corticothalamic networks during spontaneous absence seizures in rats. *J Neurosci* 22, 1480-1495.
- Mjolsness,E., DeCoste,D., 2001. Machine Learning for Science: State of the Art and Future Prospects. *Science* 293, 2051-2055.
- Moraes, M. F. 1998. *Epilepsia Experimental: estudos eletrofisiológicos e comportamentais em modelos animais de crises convulsivas audiogênicas.*
Ref Type: Thesis/Dissertation
- Morimoto,K., Fahnstock,M., Racine,R.J., 2004. Kindling and status epilepticus models of epilepsy: rewiring the brain. *Prog Neurobiol* 73, 1-60.
- Niedermeyer,E., Laws,E.R., Jr., Walker,E.A., 1969. Depth EEG findings in epileptics with generalized spike-wave complexes. *Arch Neurol* 21, 51-58.
- Paré,D., Collins,D.R., Pelletier,J.G., 2002. Amygdala oscillations and the consolidation of emotional memories. *Trends in Cognitive Sciences* 6, 306-314.
- Parra,J., Kalitzin,S.N., Lopes da Silva,F.H., 2005. Photosensitivity and visually induced seizures. *Curr Opin Neurol* 18, 155-159.
- Racine,R., Okujava,V., Chipashvili,S., 1972. Modification of seizure activity by electrical stimulation. 3. Mechanisms. *Electroencephalogr Clin Neurophysiol* 32, 295-299.
- Rodrigues,M.C., Rossetti,F., Foresti,M.L., Arisi,G.M., Furtado,M.A., Dal Col,M.L., Berti,P., Fernandes,A., Santos,F.L., Del Vecchio,F., Garcia-Cairasco,N., 2005. Correlation between shaking behaviors and seizure severity in five animal models of convulsive seizures. *Epilepsy Behav* 6, 328-336.
- Romcy-Pereira,R.N., de Araujo,D.B., Leite,J.P., Garcia-Cairasco,N., 2008. A semi-automated algorithm for studying neuronal oscillatory patterns: a wavelet-based time frequency and coherence analysis. *J Neurosci Methods* 167, 384-392.
- Romcy-Pereira,R.N., Garcia-Cairasco,N., 2003. Hippocampal cell proliferation and epileptogenesis after audiogenic kindling are not accompanied by mossy fiber sprouting or Fluoro-Jade staining. *Neuroscience* 119, 533-546.

- Rossetti,F., Rodrigues,M.C., de Oliveira,J.A., Garcia-Cairasco,N., 2006. EEG wavelet analyses of the striatum-substantia nigra pars reticulata-superior colliculus circuitry: audiogenic seizures and anticonvulsant drug administration in Wistar audiogenic rats (War strain). *Epilepsy Res* 72, 192-208.
- Sakamoto,A.C., Garzon,E., Fernandes,R.M.F., 2000. VIII EEG nas epilepsias e síndromes epilépticas. *Epilepsia* 3, 419.
- Salzberg,S.L., 1995. Locating Protein Coding Regions in Human DNA using a Decision Tree Algorithm. *Journal of Computational Biology* 2, 473-485.
- Schubert,M., Siegmund,H., Pape,H.C., Albrecht,D., 2005. Kindling-induced changes in plasticity of the rat amygdala and hippocampus. *Learn Mem* 12, 520-526.
- Seyfarth,E.A., 2006. Julius Bernstein (1839-1917): pioneer neurobiologist and biophysicist. *Biol Cybern* 94, 2-8.
- Steriade,M., 1993. Sleep oscillations in corticothalamic neuronal networks and their development into self-sustained paroxysmal activity. *Rom J Neurol Psychiatry* 31, 151-161.
- Tanabe,L., Scherf,U., Smith,L.H., Lee,J.K., Hunter,L., Weinstein,J.N., 1999. Medminer: an internet text-mining tool for biomedical information, with application to gene expression profiling. *Biotechniques* 27, 1210-1217.
- Tilelli,C.Q., Del Vecchio,F., Fernandes,A., Garcia-Cairasco,N., 2005. Different types of status epilepticus lead to different levels of brain damage in rats. *Epilepsy Behav.*
- Torrence, C. and Compo, G. P. 1998. A practical guide to wavelet analysis. *Bull.Am.Meteorol.Soc.* 79, 61-78.
- Vergnes,M., Marescaux,C., Depaulis,A., Micheletti,G., Warter,J.M., 1987. Spontaneous spike and wave discharges in thalamus and cortex in a rat model of genetic petit mal-like seizures. *Exp Neurol* 96, 127-136.
- Weiss,S.M., Indurkha,N., 1996. *Predictive Data Mining: A Pratical Guide.*
- Witten,I.H., Frank,E., 2005. *Data Mining: Practical Machine Learning Tools and Techniques.* 2nd.

ANEXOS

ANEXO 2

Algoritmo: J48 - minNumObj: 01 - Poda com fator de confiança: 0,10
 Dataset: 07 - Taxa de Crescimento e Decrescimento do Potencial e
 Freqüência no sinal de EEG baseado na média dos 05 maiores valores.



ANEXO 3

Algoritmo: PART – Regras de Produção: 07	
Dataset 07: Taxa de Crescimento e Decrescimento de Potencial Elétrico e Frequência no sinal de EEG baseado na média de 03 valores por segundo.	
Classificação	Sequência de Regras
Pré-Ictal 26 de 124 exemplos	Diminuição de Frequência em no máximo 60% no intervalo [08s, 09s] Aumento de Potencial Elétrico em no mínimo 38% no intervalo [08s, 09s] Diminuição de Frequência em no máximo 61% no intervalo [09s, 10s]
Pré-Ictal 16 de 124 exemplos	Aumento de Potencial Elétrico em no máximo 30% no intervalo [01s, 02s] Diminuição de Potencial Elétrico em no máximo 17% no intervalo [03s, 04s] Aumento de Frequência em no máximo 11% no intervalo [06s, 07s] Aumento de Potencial Elétrico entre 1.000% e 3.000% no intervalo [09s, 10s]
Pré-Ictal 20 de 124 exemplos	Aumento de Potencial Elétrico entre 1.000% e 4.000% no intervalo [09s, 10s]
Fim-Ictal 22 de 124 exemplos	Aumento de Potencial Elétrico entre 01% a 50% no intervalo [02s, 03s] Diminuição de Frequência entre 07% a 44% no intervalo [06s, 07s] Diminuição de Potencial Elétrico entre 01% a 90% no intervalo [07s, 08s] Diminuição de Frequência entre 01% a 60% no intervalo [08s, 09s]
Fim-Ictal 20 de 124 exemplos	Aumento de Potencial Elétrico entre 01% a 50% no intervalo [02s, 03s] Aumento de Potencial Elétrico em no mínimo 20% no intervalo [08s, 09s]
Fim-Ictal 18 de 124 exemplos	Aumento de Frequência em no máximo 27% no intervalo [01s, 02s] Diminuição de Frequência entre 01% a 22% no intervalo [02s, 03s] Diminuição de Potencial Elétrico entre 01% a 28% no intervalo [08s, 09s]
Fim-Ictal 02 de 124 exemplos	Não foram cobertos por nenhuma regra.

ANEXO 4

Algoritmo: PART – Regras de Produção: 05	
Dataset 07: Taxa de Crescimento e Decrescimento de Potencial Elétrico e Frequência no sinal de EEG baseado na média de 05 valores por segundo.	
Classificação	Sequência de Regras
Pré-Ictal 33 de 124 exemplos	Média de Potencial Elétrico de 0,069 mV no tempo 01s. Diminuição de Potencial Elétrico em no máximo 15% no intervalo [06s, 07s] Aumento de Potencial Elétrico em no máximo 500% no intervalo [09s, 10s]
Pré-Ictal 29 de 124 exemplos	Aumento de Potencial Elétrico em no máximo 538% no intervalo [08s, 09s]
Fim-Ictal 23 de 124 exemplos	Diminuição de Frequência em no máximo 23% no intervalo [03s, 04s] Diminuição de Frequência em no máximo 29% no intervalo [07s, 08s] Aumento de Potencial Elétrico em no máximo 530% no intervalo [08s, 09s] Diminuição de Potencial Elétrico entre 01% e 43% no intervalo [09s, 10s]
Fim-Ictal 24 de 124 exemplos	Aumento de Potencial Elétrico em no máximo 660% no intervalo [02s, 03s] Aumento de Potencial Elétrico em no mínimo 01% no intervalo [09s, 10s]
Fim-Ictal 15 de 124 exemplos	Não foram cobertos por nenhuma regra.

ANEXO 5

1:	Iniciar processo de corte em sinais digitais
2:	
3:	Solicitar entrada do(s) arquivo(s) de sinal(is)
4:	Solicitar valor da Frequência de Amostragem – ‘Fs’
5:	Solicitar o tempo do trecho a ser cortado – ‘tempo’
6:	Pontos = tempo*Fs
7:	
8:	Repita: Para cada arquivo inserido faça
9:	Solicitar tempo inicial do corte – ‘TempoInicial’
10:	Solicitar tempo final do corte – ‘TempoFinal’
11:	
12:	Arquivo Pré-ictal = ArquivoOriginal ((TempoInicial – 10)*Fs , TempoInicial*Fs)
13:	Arquivo Fim-ictal = ArquivoOriginal ((TempoFinal – 10)*Fs , TempoFinal*Fs)
14:	Fim_Repita
15:	
16:	Fim_Rotina

Algoritmo desenvolvido para separar épocas de interesse dentro de um sinal de EEG.

ANEXO 6

1:	Iniciar Rotina de aplicação da Transformada de Fourier
2:	
3:	Solicitar entrada do(s) arquivo(s) de sinal(is)
4:	Solicitar valor da Frequência de Amostragem – ‘Fs’
5:	Solicitar o tempo do trecho a ser transformado – ‘tempo’
6:	Freq = [0, Fs/2]
7:	
8:	Repita: Para cada arquivo inserido faça
9:	Plotar em gráfico sinal do EEG vs. Tempo
10:	Repita: Para cada segundo de 0 até Tempo
11:	Transformada Fourier no sinal – FFT(sinal)
12:	Extração da parte real do resultado (Amplitude)
13:	Extração dos 05 maiores valores de potência e Frequência da FFT
14:	Plotagem da Amplitude do sinal analisado
15:	Salva figura de gráficos e arquivo texto de potências e frequências
16:	Fim_Repita
17:	
18:	Fim_Repita
19:	
20:	Fim_Rotina

Algoritmo desenvolvido para extrair o Domínio Espectral do sinal digital através da Transformada de Fourier

ANEXO 7

1:	Iniciar Rotina de aplicação da Transformada de Wavelet
2:	
3:	Solicitar entrada do(s) arquivo(s) de sinal(is)
4:	Solicitar valor da Frequência de Amostragem – ‘Fs’
5:	Solicitar o tempo do trecho a ser transformado – ‘tempo’
6:	Solicitar Vetor de Frequencias – ‘freqVec’
7:	Solicitar tamanho da janela de análise (default: 5)
8:	Freq = [0, Fs/2]
9:	TimeVec = [1, (Fs*tempo)] / Fs
10:	Repita: Para cada posição no sinal amostrado (Fs*tempo vezes)
11:	Repita: Para cada posição no vetor de frequencias
12:	Matriz = Valor médio da energia obtida pelas tentativas de janelamento pela Wavelet de Morlet.
13:	Fim_Repita
14:	Fim_Repita
15:	Plotagem dos valores de energia vs. freqVec
16:	Fim_Rotina

Algoritmo desenvolvido para extrair o Domínio Espectral do sinal digital através da Transformada de Wavelet