Predição do diagnóstico da epilepsia utilizando variações no número de cópias dos genes

Trabalho de Conclusão de Curso

Engenharia da Computação

Fernando Oliveira
Orientador: Prof. Meuser Valença

Universidade de Pernambuco
Escola Politécnica de Pernambuco
Graduação em Engenharia de Computação

Fernando Oliveira de Araújo Júnior

Predição do diagnóstico da epilepsia utilizando variações no número de cópias dos genes

Monografia apresentada como requisito parcial para obtenção do diploma de Bacharel em Engenharia de Computação pela Escola Politécnica de Pernambuco – Universidade de Pernambuco.

Recife, Dezembro de 2014.

De acordo

Recife

____/_____/_____

_____
Meuser Valença

Agradecimentos

Agradeço ao meu orientador Meuser Valença pela oportunidade de ser seu orientando.

Resumo

Neste trabalho de conclusão de curso são utilizados modelos de predição (*random forest*, *support vector machines* e *k-nearest neighbor*) para realizar o diagnóstico de pacientes epiléticos e não epiléticos com base em variações de cópias dos genes (*copy-number variations*). Os resultados são positivos, com aproximadamente 90% de acerto médio para o melhor modelo encontrado, o que mostra que é possivel o uso desse tipo de tecnologia no auxilio ao diagnóstico desta doença de causa ainda desconhecida.

Abstract

In this work, prediction models (random forest, support vector machines and k-nearest neighbor) are used to perform the diagnostic of epileptic and non-epileptic patients using copy-number variations of genes. The results are positive, with approximately 90% of accuracy for the best model found, which shows that is possible to use this type of technology as a support in the diagnostic of this disease whose the cause is still unknown.

Sumário

Índice de Figuras

Tabelas de Símbolos e Siglas

AED – Anti-Epileptic Drugs (Drogas anti-epiléticas)

CGH – Comparative Genomic Hibridization (Hibridização Genômica Comparativa)

CNV – Copy Number Variation (Variação no Número de Cópias)

DNA – Deoxyribonucleic acid (Ácido Desoxirribonucleico)

HIV – Human Immunodeficiency Virus (Vírus da Imunodeficiência Humana)

IQR – Interquartile Range (Distância entre quartis)

KNN – K-Nearest Neighbor (K Vizinhos mais Próximos)

ROC – Receiver Operating Characteristic (Característica de Recepção do Operador)

SNP – Single Nucleotide Polymorphism (Polimorfismo de Nucleotídeo Único)

SVM – Support Vector Machines (Máquinas de Vetores de Suporte)

Introdução

A epilepsia é um distúrbio neurológico até o momento sem cura e, tratando-se de epilepsia não causada por traumas (pancada na cabeça, infecção cerebral, tumores), sem conhecimento de causa ou origem, .

Seu tratamento é feito através de remédios controlados conhecidos como AEDs (*Anti-Epileptic Drugs*) que tem como foco o controle das convulsões. Eles causam muitos efeitos colaterais desagradáveis em seus pacientes, além de não agir diretamente no foco da doença, o que atualmente não é possível.

A caracterização do distúrbio, ou seja, a descoberta dos fatores genéticos e proteicos que contribuem diretamente para a sua existência, é essencial para que se chegue a um tratamento mais eficaz.

1.1 Motivação e Caracaterização do Problema

Aproximadamente 1% da população mundial sofre de epilepsia. Apesar de boa parte dos pacientes viverem quase que normalmente, esta doença traz muitos transtornos para boa parte dos epiléticos, em alguns casos levando-os a óbito.

Além disso, a epilepsia faz parte de um grupo de doenças (autismo, esquisofrenia, transtorno bipolar, retardamento mental) que aparentemente têm origem em variações genéticas que são difíceis de identificar através dos testes convencionais feitos em laboratórios.

O entendimento da causa/origem genética destas doenças é importante para o avanço da medicina e para o entendimento do genoma humano.

1.2 Hipóteses e Objetivos

A epilepsia, quando não causada por trauma, é causada por mutações genéticas. O diagnóstico desta doença é feito por meio de alguns exames, nem sempre conclusivos, após o paciente ter sofrido no mínimo uma convulsão.

Este trabalho tem como principal objetivo diagnosticar pacientes epiléticos através de amostras de variações genéticas em seu DNA.

1.3 Organização do Documento

Este trabalho está organizado em 5 capítulos. O Capítulo 2 apresenta conceitos sobre genética e contextualiza alguns dos aspectos que motivam este trabalho: a epilepsia e variações genéticas. O Capítulo 3 aborda as ferramentas que fazem este trabalho produzir resultados, técnias de seleção, modelos de predição, *machine learning*. O Capítulo 4 é um resumo do experimento, como o banco foi minerado e como seus dados foram utilizados pelos modelos de predição. O Capítulo 5 conclui o trabalho fazendo uma análise do que foi feito e dos resultados, bem como possíveis trabalhos futuros.

Epilepsia e Variação Genética

2.1 Gene

São seções de uma molécula de ácido desoxirribonucleico (DNA) que, por sua vez, é responsável por carregar toda informação hereditária dos seres vivos.

A molécula de DNA é composta por uma "fita" de dupla hélice de pares de bases nitrogenadas que se ligam por pontes de hidrogênio. São elas: adenina (A), timina (T), guanina (G) e citosina (C), conforme mostrado na Figura 1.

Os genes são responsáveis pela produção das proteínas. Cada gene produz uma proteína diferente. Elas são as principais macromoléculas do organismo. A sequência do amino ácido presente em seu interior é codificada por algum gene. O tempo e taxa de produção das proteínas e outros componentes celulares são funções de ambos: os genes e o ambiente externo ao organismo [2].

Cada gene pode ser encontrado em várias formas diferentes, geralmente em pequenas formas. Essas formas são chamadas alelos. Variação alélica causa variações hereditárias dentro de uma espécie. Em nível protéico, variação alélica se torna variação protéica.

Cada cadeia de nucleotídeos é ligada pelas extremidades entre o açúcar e as porções de fosfato dos consecutivos nucleotideos, isto é chamado de “esqueleto da cadeia”. As duas cadeias entrelaçadas são mantidas juntas por uma ligação fraca entre as bases em “fios” opostos onde, adenina parea apenas com timina e guanina parea apenas com citosina. As bases que paream são ditas complementares.

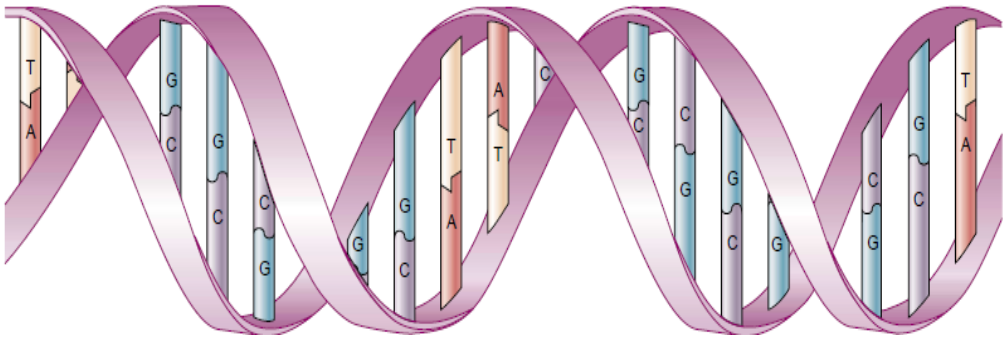

Figura 1. Estrutura do DNA humano.

Uma proteína geralmente tem uma de duas funções [2], dependendo de seu gene. A primeira função é a de componente estrutural, contribuindo para propriedades físicas das células ou organismos, como por exemplo: microtubulos, músculo, proteínas do cabelo. A segunda função é a de agente ativadora em processos celulares, assim como enzimas que catalizam alguma reação química da célula.

2.2 Cromossomo

Cromossomos são as estruturas dentro do núcleo das células que carregam o DNA. Eles existem em pares (eucariontes) ou sozinhos (procariontes) e são enumerados de 1 a 22, aproximadamente, do mais largo para o menos largo. Nos seres humanos existem ainda dois cromossomos X e Y, chamados cromossomos do sexo. Mulheres possuem dois cromossomos Xs (XX) e homens possuem um X e um Y (XY).

O cromossomo pode ser dividido em 3 partes: braço curto, centrômero e braço longo. Ao longo desses braços estão localizados os genes (alelos).

2.3 Variação Genética

O DNA humano codifica aproximadamente 30 mil genes. Cada ser humano possui duas cópias de cada gene, assim como possui um par de cromossomos. Porém, existem variações com relação ao número dessas cópias (*copy-number variation ou CNV*). Essa variação é definida como a deleção ou a duplicação de

pedaços de DNA maiores ou iguais a 1kb [8] (kilobase, mil pares de bases nitrogenadas) e que geralmente englobam um ou mais genes [10], conforme a Figura 2.

Diferenças na sequência do DNA do genoma humano contribuem fortemente para a unicidade de cada humem ou mulher. As CNVs influenciam muitas características incluindo a suscetibilidade a doenças. Estudos recentes revelam que CNVs abrangem pelo menos 3 vezes o número total de conteudo, em nucleotídeos, dos SNPs (*Single Nucleotide Polymorphism*), que até pouco tempo eram tidos como a forma de variação genética mais significativa para nossa diferenciação.

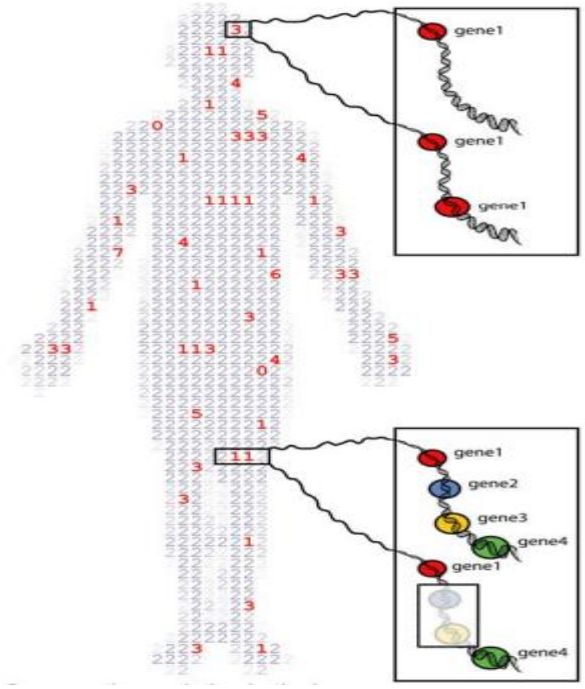

Figura 2. Variações de cópias dos genes, duplicação e deleção.

Alguns estudos tem demonstrado a importância das CNVs como variantes de suscetibilidade a doenças [4], com específicas CNVs identificadas como associadas a mudança no risco de infecções HIV, doenças auto-imune, e asma. Recentemente, estudos no genoma humano completo tem demonstrado que raras CNVs alteram genes, que são, na verdade ligados ao desenvolvimento neural, implicando em desordens do espectro autista e esquizofrênia.

2.4 ***Comparative Genomic Hibridization***

Alterações no material cromossômico, ganho (duplicação de nucleotideos) ou perda (deleção de nucleotideos) podem ser detectados por testes clínicos analíticos. Alguns testes de cromossomos feitos por cientistas clínicos não conseguem detectar ganhos ou perdas muito pequenas. Uma abordagem para este tipo de problema é o teste CGH (hibridização genômica comparativa) baseado em *microarrays*.

Arrays CGH comparam o DNA do paciente doente com uma amostra de DNA de alguém identificado como “controlado” ou “não doente”, chamado DNA de referência. Desta forma, variações/mutações (deleções e duplicações) entre os dois conjuntos de DNA são estabelecidas. Os resultados dessas comparações são relativos, pois dependem do DNA de referência. E ainda, quanto maior for a duplicação, maior será seu valor atribuído com sinal positivo, por outro lado, quanto maior for a deleção, maior será seu valor atribuído com sinal negativo.

O funcionamento do teste CGH se dá através da capacidade da molécula de DNA (que tem formato de fio) de se ligar/parear a uma outra molécula de DNA específica que depende da sequência de nucleotídeos. Este fenônemo é chamado de hibridização.

Cada *microarray* abrange milhares de pequenas sequências de DNA (“probes”), arrumadas em uma grade em uma lâmina de vidro. O DNA do paciente é “digerido”, fatiado em pequenos fios (ou fragmentos). Esses fragmentos são rotulados com um corante de cor fluorescente. O DNA de referência é rotulado com um coloração fluorescente diferente. As duas amostras são misturadas e aplicadas em uma lâmina de vidro onde a hibridização começa, ou seja, os DNAs se paream com seus “probes” correspondentes. A lâmina então passa por um “*microarray scanner*” que mede a quantidade de fluorescência de cada cor em cada probe pareado, conforme a Figura 3.

Os dois tipos de probes mais comumente utilizados por *array* CGH para detecção de mudanças cromossômicas são *bacterial artificial chromosome* (BAC) e oligonucleotide (oligo). A principal diferença entre os dois probes resultantes está principalmente em seu tamanho. BACs que possuem geralmente de 80 a 200 kbs

(kilobases) tendem a não detectar pequenas variações. Já Oligos são probes muito menores, por volta de 60 bps (base-pair) [11].

A principal vantagem de arrays CGH é a possibilidade de explorar todos os 46 cromossomos (em caso de humanos) em um único teste. Porém ele não capta todo tipo de variações. Inversões e translocações (que não resultam em ganho ou perda do material chromossomico) não serão identificadas.

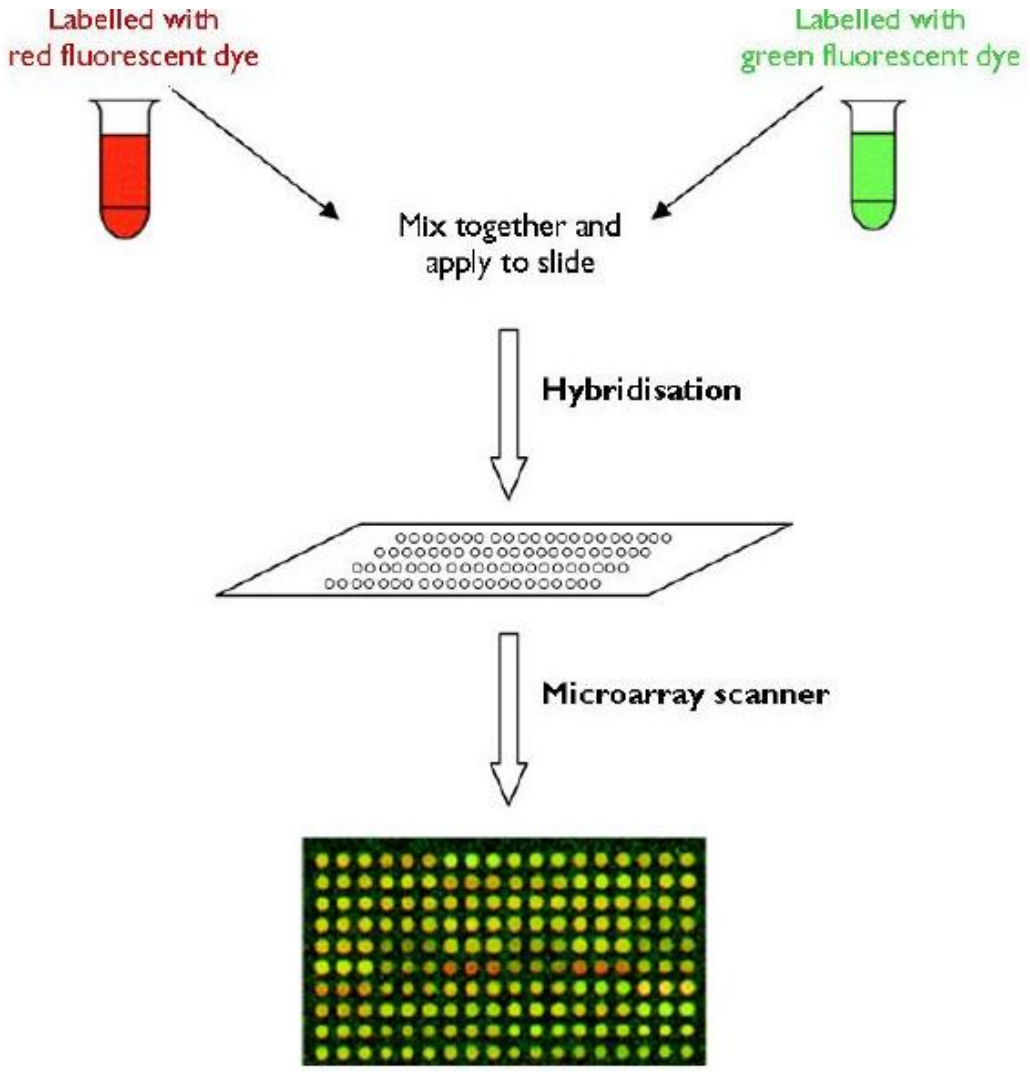

Figura 3. Teste de CGH.

Data Mining

Data mining consiste de técnicas e algoritmos que permitem visões fundamentais e conhecimento de grande massa de dados. É um campo que mistura conceitos de áreas correlacionadas como sistemas de bancos de dados, estatística, *machine learning* e reconhecimento de padrões. Sua aplicação é um amplo processo de descoberta de conhecimento, que inclui tarefas de pré-processamento como extração de dados, limpeza de dados, fusão de dados, redução de dados, construção de características, e também etapas de pós-processamento como interpretação de modelos e padrões, geração e confirmação de hipóteses. Essa descoberta do conhecimento tende a ser um processo altamente iterativo e interativo [17].

3.1 Seleção de Características

Seleção de características é uma tarefa importante em muitos problemas de *data mining*. A questão geral é selecionar um conjunto de característias (variáveis) do problema que seja mais relevante para a análise dos dados [1].

No geral, existem dois tipos de abordagens para a seleção de características: filtros e empacotadores. Filtros acontecem em um único passo, enquanto empacotadores geralmente envolvem um processo de busca onde a cada ciclo se procura por um conjunto de características que seja mais adequado para as ferramentas de *data mining* que serão aplicadas. Por isso empacotadores consomem mais recursos computacionais, o que os torna menos adequados para problemas com grandes quantidades de dados. Eles costumam selecionar características que sejam mais adequadas para as ferramentas que serão utilizadas. Neste trabalho são utilizados filtros para a seleção de características.

3.1.1 Seleção Baseada em Propriedades de Distribuição

Este filtro é baseado na informação de distribuição contida em cada característica, neste caso, o valor em módulo das variações por número de cópias

(*copy-number variations*). Isto é, uma característica que varia pouco entre as amostras dificilmente servirá para diferenciar uma amostra de outra.

Para representar a distribuição das CNVs, neste trabalho é utilizada a medida de distância entre-quartis (IQR), onde os quartis são 3 divisões: o primero quartil representa 25% dos dados, os que possuem menor expressão, ou as mais negativas; o segundo quartil é chamado também de mediana e representa um corte que divide metade dos dados a sua esquerda e metade a sua direita; o terceiro quartil representa os últimos 25% dos dados e concentra os valores mais expressivos positivos. O valor de IQR é dado pela diferença entre o terceiro quartil e o primeiro, representando assim uma medida de variabilidade.

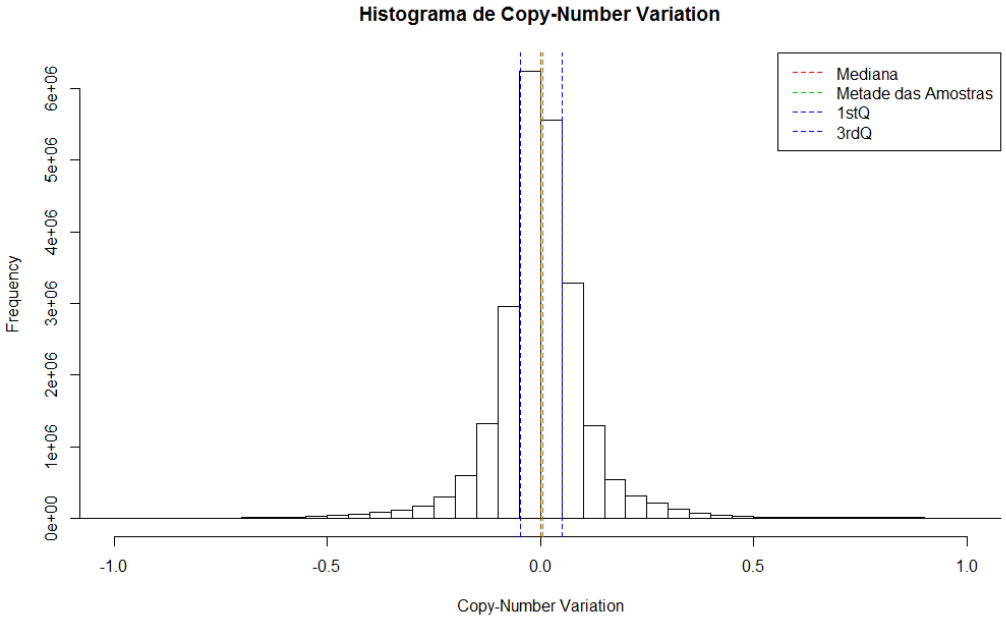

Figura 4. Histograma de distribuição das caracaterísticas (expressão x frequência)

A Figura 4 representa o histograma de expressão das características (CNVs) do banco utilizado. Ele mostra que, em média, a grande maioria das características tem expressividade próxima ou igual a zero.

O objetivo desse filtro é remover características com pouca variabilidade, o que neste caso representa a maior parte das características.

3.1.2 Seleção Baseada em Análise de Variância (ANOVA)

Análise de Variância (ANOVA) é uma técnica que analisa fatores (neste trabalho são as características) que variam quando comparados grupo a grupo. O processo consiste em definir um grupo de fatores de tamanho N para cada classe (neste trabalho são os três possíveis diagnósticos) e calcular a variância (v) para todos os elementos dos grupos (numero de grupos x N) utilizando a seguinte fórmula:

$$v = \sigma/Q^{½}$$

onde, σ é o desvio padrão e N é a quantidade total de elementos.

A análise é feita com resultados de vários testes que geram como resultado o fator F dado por:

F = (desvio padrão calculado) / (desvio padrão esperado)

onde, o numerador dado pelo desvio padrão de todos os elementos que foram selecionados para o teste corrente e o denominador é dado pela fórmula da variação (v). A seleção do grupo a ser testado é feita mantendo o grupo da classe a ser testada intacto e alterando aleatoriamente os elementos dos grupos de outras classes. Se F for aproximadamente 1 na maioria dos casos, então a variância não é considerável.

3.1.3 Seleção utilizando Árvores de Decisão

Aprendizado por árvore de decisão é um método que consiste em aproximar funções discretas que são robustas a dados ruidosos e que são capazes de aprender expressões disjuntas [19]. Neste caso a função é representada por uma árvore de decisão.

Árvores de decisão podem ser representadas por conjuntos de regras 'if-else' e estão entre os algoritmos de inferência indutiva mais populares e tem sido aplicado com sucesso em muitos problemas que vão de diagnósticos médicos a avaliações de risco de empréstimos para aplicações [19].

Árvores de decisão classificam instâncias ordenando-as de cima para baixo na árvores começando pela raíz até alguma folha que fornece a classificação da instância. Cada nó da árvore especifica um teste de algum atributo da instância

enquanto cada galho descendente daquele nó corresponde a um dos possíveis valores para aquele atributo. Uma instância é classificada começando pelo nó raíz da árvore, testando o atributo especificado para este nó e então movendo para baixo pelo galho correspondente até o próximo nó ou folha (fim do processo).

Problemas apropriados para Aprendizado por Árvores de Decisão são os seguintes:

- Instâncias são representadas por valores em pares: (sim ou não, quente ou frio, etc...);
- A função alvo é composta por valores discretos: a grande maioria dos problemas que utilizam árvores de decisão é representado por funções discretas. Existem ainda extensões que permitem a utilização deste método, menos usual, em problemas com funções reais ou contínuas;
- Descrições disjuntivas podem ser requeridas: árvores de decisão naturalmente representam expressões disjuntivas;
- O conjunto de dados de treinamento pode conter erros: são robustas tanto com relação a erros de classificação quanto a erros nos atributos;
- O conjunto de dados pode apresentar falta de valores: são robustas quando atributos apresentam valores desconhecidos.

Um conceito muito importante no contexto das Árvores de Decisão é a Entropia. Ela mede a homogeneidade das amostras, caracterizada pela impureza de uma coleção arbritária de amostras. Se a entropia se aproxima de 0,5 então o conjunto de amostras está homogênio.

Dada uma coleção S, contendo amostras positivas e negativas de algum conceito alvo, a entropia de S relativa à essa classificação booleana é dada por:

Entropia(S) = $-\rho_{\oplus} \log_2 \rho_{\oplus} - \rho_{\ominus} \log_2 \rho_{\ominus}$

onde $\rho_{\oplus}$ é a proporção positiva de amostras em S e $\rho_{\ominus}$ é a proporção negativa de amostras em S e ($0 \log_2 0$) é definido como 0.

Como exemplo, dada uma coleção S de 14 amostras de algum conceito booleano (possui apenas dois valores possíveis) onde 9 são positivas e 5 são negativas, a entropia de S relativa à classificação é dada por:

Entropia(9+,5-) = -(9/14)$\log_2(9/14)$ - (5/14) $\log_2(5/14)$ = 0.940

Mais genericamente, se uma instância alvo pode ser classificada por c rótulos diferentes, então a entropia S relativa a essa classificação é definida por:

Entropia(S) = $\sum_{i=1}^{c} -\rho i \log_2 pi$

onde ρi é a proporção de S pertencente a classe c(i).

Para que a classificação ocorra mais rápido, é necessário que os classificadores (atributos) mais eficientes sejam ordenados do topo da (raiz) árvore para baixo. Essa ordenação é feita através do ganho de informação de cada atributo, que é a representação da perda de entropia que o conjunto de amostras sofre quando é particionado pelo atributo em questão. O ganho do conjunto S particionado pelo atributo A é dado por:

Ganho(S, A) = Entropia(S) - $\sum_{v \in Values(A)} \frac{|Sv|}{|S|} Entropia(S)$

onde S é o conjunto de amostras, Values(A) é o conjunto de valores que o atributo A pode receber, Sv é o subconjunto de amostras onde o atributo A recebe o valor v.

A seleção por árvores de decisão ocorre definindo-se as características que tem maior ganho.

3.2 Modelos de Predição

3.2.1 Random Forest

Random forest é um modelo constituído por um conjunto de modelos mais simples, as árvores de decisão. Este modelo pode ser usado tanto em problemas de classificação quanto para problemas de regressão e ainda em problemas de seleção de características.

O número de árvores no modelo é configurado pelo usuário. As árvores são constituidas de características escolhidas aleatoriamente do conjunto de dados, ou seja, cada árvore tem um aprendizado diferente e geram resultados independentes das outras árvores.

A predição do conjunto total é obitida pela contagem de votação de todas as árvores.

3.2.2 K-nearest neighbor

K-nearest neighbors é um algoritmo que pertence a classe dos chamados "*lazy learners*" pois não obtem um modelo através de treinamento, ele apenas registram o conjunto de dados (amostras).

Na forma mais pura do algoritmo, a predição ou classificação de uma amostra é definida pelos k vizinhos mais próximos a ela, ou seja, se a maioria dos k vizinhos pertence a uma classe X, então a amostra também será classificada como da mesma classe, conforme a Figura 5. Existem ainda extensões que dão alguma prioridade às amostras mais próximas dentre os k vizinhos escolhidos, criando assim uma votação que pondera classe e distância na predição da amostra alvo.

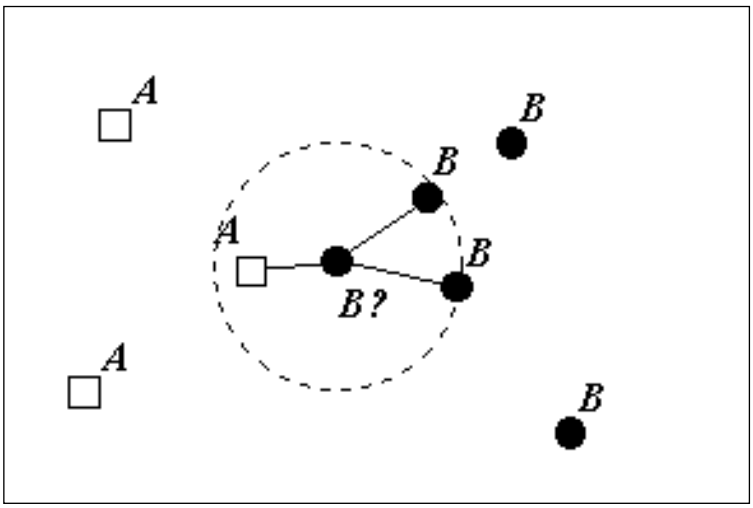

Figura 5. Modelo KNN com as amostras representadas num plano.

3.2.3 Support Vector Machines

Para desenvolver um modelo computacional de aprendizado supervisionado, deve-se escolher uma forma de representar funções/hipóteses no computador. Usualmente, utiliza-se uma função linear de x para uma aproximação de y, tal como: $h(x) = w_0x_0 + w_1x_1 + w_2x_2$ [18], onde w_i são parâmetros (também chamados de *weights*) que parametrizam o espaço linear criando um mapeamento de x para y.

$h(x) = \sum_{i=0}^{N} wixi = g(w^Tx)$ [18]

SVM é um modelo de aprendizado supervisionado onde os parâmetros (*weights*) representam um vetor ortogonal a um hiperplano que separa os grupos de dados que são divididos por sua classificação, conforme a Figura 6.

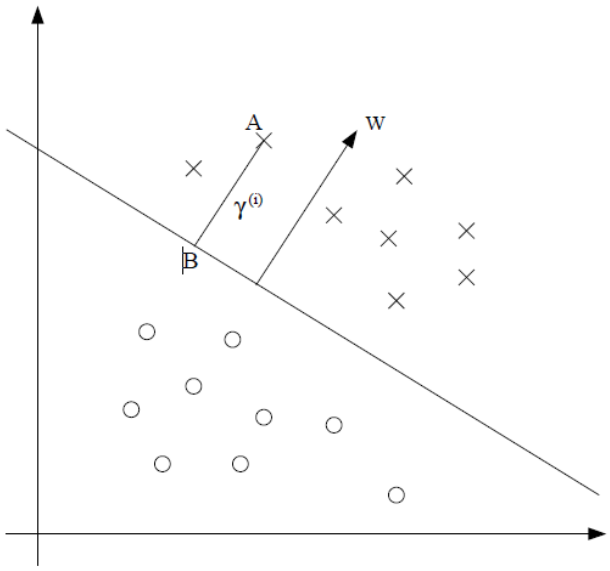

Figura 6. SVM representanda num plano

Para um conjunto de amostras classificadas com rótulos binários y e características x, tem-se y ϵ {-1,1} que denotam os dois valores possíveis de classificação da amostra. Além disso, a função/ hipótese h é definida através de dois parâmetros w (*weight*) e uma constante b através da seguinte fórmula

$h_{w,b}(x) = g(w^Tx+ b)$, onde:

$$h_{w,b}(x) = \begin{cases} 1 \; se \; g \geq 0 \\ -1 \; caso \; contrário \end{cases}$$

Margem (m) é a menor distância de uma amostra X ao hiperplano que separa dois conjuntos de dados [18]. Esse hiperplano é a reta dada por $h_{w,b}(x) = 0$ (limiar de decisão, classificação).

$m = \min(m_i)$, onde

$m_i = y(i) * h_{w,b}(x)$ (margem funcional)

Neste caso, se y(i) = 1, então o melhor caso de $g(w^T x + b)$ é um valor grande em módulo e positivo. Por outro lado, se y(i) = -1, então para que a margem seja larga, $g(w^T x + b)$ tem de ser um valor grande em módulo e de sinal negativo. Pode-se ainda dizer que se m(i) > 0, então a predição está correta e quanto maior a margem maior a confiabilidade do modelo.

Geometricamente, a distância da amostra A ao hiperplano $\gamma(i)$ é dada pelo seguimento AB. Tem-se também que $\frac{w}{||w||}$ é um vetor unitario na direção de $\overline{AB}$. Pode-se dizer então que $A-B = \gamma^{(i)} \frac{w}{||w||}$. Logo $B = A - \gamma^{(i)} \frac{w}{||w||}$, e como h(B) = 0:

$$w^T \left(x^{(i)} - \gamma^{(i)} \frac{w}{||w||}\right) + b = 0.$$

Resolvendo para $\gamma^{(i)}$:

$$\gamma^{(i)} = \frac{wTx(i) + b}{||w||} = \left(\frac{w}{||w||}\right)^T x^{(i)} + \frac{b}{||w||}.$$ [18]

Generalizando para os dois conjuntos (positivo e negativo)

$$\gamma^{(i)} = y^{(i)} * \left(\left(\frac{w}{||w||}\right)^T x^{(i)} + \frac{b}{||w||}\right)$$ (margem geométrica)

Pode-se notar que se ||w|| = 1, então a margem funcional é igual a margem geométrica. Além disso, a magnitude do vetor w e a contante b podem variar indefinidamente para se ajustarem ao conjunto de treinamento[18].

A margem geométrica também é definida como a menor margem das amostras:

$m = \min(m_i)$

Para a determinação do melhor w, são utilizados algorítimos de otimização quadrática ou ainda técnicas matemáticas como Dualidade de Lagrange. Ao se obter o w, a SVM estará pronta para ser utilizada.

3.3 Métrica de Avaliação

A tarefa de predição é um problema de classificação multi-classe [1]. Modelos de classificação preditiva são geralmente avaliados utilizando a taxa de erro ou o seu complemento, o acerto. Contudo, existem várias outras alternativas, como área sobre a curva ROC [23] e medidas de precisão de estimativa de probabilidade de classe (*Brier Score*) [24].

A seleção para a medida de avaliação para um dado problema depende dos objetivos a serem alcançados e das características do banco de dados, como por exemplo o custo do erro de uma classificação incorreta. Neste trabalho é considerado que qualquer erro tem o mesmo custo.

Em problemas com mais de duas classes, análises de curva ROC não são muito bem estabelecidas [1]. Neste contexto, este trabalho usa a medida padrão de precisão como métrica de avaliação para os resultados gerados, e é dada pela seguinte fórmula:

$$\overline{acc} = 1 - \frac{1}{N} \sum_{i=1}^{N} L_{0/1}(yi, \hat{y}i),$$

onde N é o número de elementos (amostras), e $L_{0/1}()$ é uma função definida como:

$$L_{0/1}(yi, \hat{y}i) = \begin{cases} 0 \; se \; yi = \hat{y}i \\ 1 \; caso \; contrário \end{cases}$$

3.4 Validação Cruzada

Validação cruzada é um método estatístico de avaliação e comparação de algoritmos de aprendizado que consiste em dividir os dados em dois segmentos: o primeiro, chamado de conjunto de treinamento, é utilizado na formação do modelo e

em seu aprendizado, e o segundo é utilizado para avaliar a efetividade do modelo utilizando alguma métrica de avaliação para realizar a medida, conforme a Figura 7.

A forma mais convencional de validação cruzada é chamada k-fold, mas existem outras formas, que são casos especiais da forma convencional ou envolvem repetidas iterações [20].

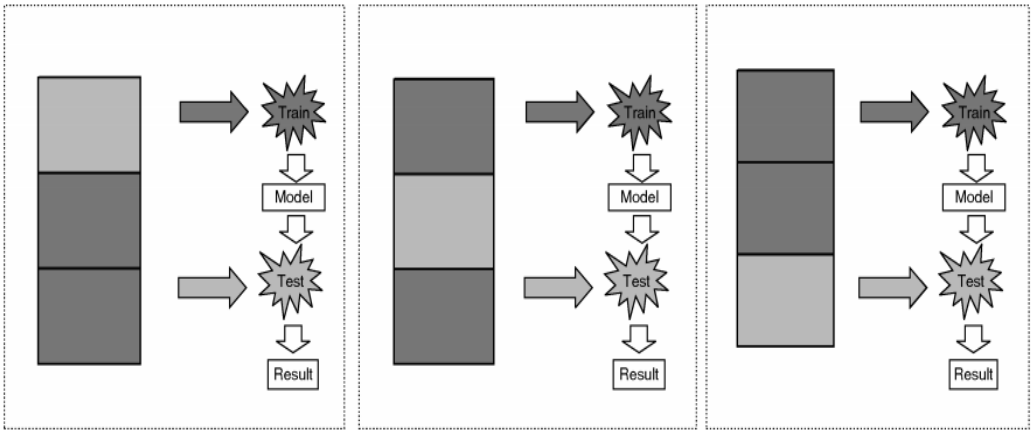

Figura 7. Exemplo do fluxo de formação e treinamento dos modelos, bem como a geração de resultados.

Existem dois objetivos principais quando se usa a validação cruzada:

- Estimar o desempenho de aprendizagem do modelo sobre dados disponíveis usando algum algoritmo, isto é, medir sua capacidade de generalização.
- Para comparar o desempenho de dois ou mais algoritmos e descobrir qual o melhor para um determinado conjunto de dados, ou alternativamente, para comparar o desempenho de duas variantes de um modelo parametrizado.

3.4.1 Hold-Out Validation

Esta técnica de validação cruzada é focada em evitar *over-fitting* e tem como principal característica a escolha de um conjunto de testes independente. A abordagem natural para esta técnica é separar os dados em duas partes distintas: uma para treino e outra para teste. Os dados para teste não são acessados durante o treinamento. Hold-Out não permite sobreposição entre os dois conjuntos de dados,

permitindo assim uma estimativa de precisão para o desempenho do algoritmo de forma generalizada. Por outro lado, o fato de os testes serem feitos em um subconjunto do total de dados faz com que os resultados estejam atrelados aos dados selecionados para teste assim como o aprendizado será dependente dos dados selecionados para treinamento, ou seja, a separação dos dados é uma fase crítica e influencia diretamente no desempenho do modelo. Esse problema pode ser parcialmente evitado (ou contornado) com a repetição do processo múltiplas vezes e tomando a média dos resultados como resultado final da validação. Neste caso, existe ainda a possibilidade de repetir um subconjunto dos dados muitas vezes no conjunto de dados de testes, ou ainda, nunca serem utilizadas no conjunto de teste, provocando assim algum over-fitting. Para evitar este problema, a forma convencional da validação cruzada (k-fold) pode ser utilizada.

3.4.2 K-Fold Cross Validation

Na validação cruzada do tipo *k-fold* os dados são primeiramente distribuídos igualmente em k segmentos. São realizadas K iterações sequênciais de treinamento e validação onde, em cada iteração, um segmento diferente de dados é separado para a validação enquanto que os remanescentes k - 1 segmentos são usados para treinamento.

3.4.3 *Leave-OneOut Cross Validation* (LOOCV)

Leave-one-out é um caso especial da validação do tipo *k-fold* onde o número de elementos em cada segmento é igual a um, ou seja, a cada iteração um dado é utilizado para validação enquanto que todos os outros são utilizados para treinamento. Este método é especialmente utilizado para conjuntos de dados muito pequenos.

Experimento

Este capítulo explica todas as etapas da parte prática deste trabalho, desde o tratamento dos dados até a aplicação dos modelos. O tratamento dos dados é feito em 4 passos e é focado na seleção das características que são mais relevantes para que os modelos consigam diferenciar as amostras. Ao final da seleção, 30 características (CNVs) são utilizadas pelos três modelos: SVM, Random Forest e KNN, para que seja realizada a predição.

A tecnologia utilizada foi a plataforma R, que é um ambiente de software livre para estatística computacional e gráficos. A versão utilizada foi a 3.1.2 e boa parte do código foi baseado nas *packages* da Bioconductor.

4.1 Banco de Dados

O banco de dados utilizado neste trabalho contempla 483 amostras onde 235 são pacientes com duplo diagnóstico de esquizofrenia e epilepsia, 80 pacientes com duplo diagnóstico de transtorno bipolar e epilepsia e 168 pacientes com quadro psiquiátrico no estado denominado de "*control*" (controlado). Foi publicado em 18 de agosto de 2010 pelo Doutor Arthur Beaudet do Departamento de Genética Humana e Molecular do Colégio de Medicina de Baylor na cidade de Houston, e teve sua última atualização em 22 de março de 2012.

Cada amostra contem um *array* com 51183 mutações cromossômicas (CNVs) que foram obitidas por hibridização genômica comparativa (CGH) utilizando um mesmo ser humano como referência.

O banco conta ainda com informações da posição da variação no cromossômico, sequência genética, sexo das amostras e tipo de célula utilizada.

O banco é livre e está acessível na Internet [22].

4.2 Seleção das Caracaterísticas (CNV)

Essa etapa do projeto é possivelmente a mais importante pois define as 30 características que serão selecionadas dentre as 51183 existentes no banco e que serão a ferramenta base para os modelos de predição funcionarem.

O foco deste trabalho é a predição de pacientes epiléticos. Porém, como a epilepsia é uma doença em parte causada por variações genéticas, existe um grau de importânica nos genes ligados as CNVs muito grande. Muitos trabalhos nos últimos anos vem tentando encontrar um gene ou um grupo de genes ou mutações que participem diretamente na causa da epilepsia e, como este trabalho prediz, com certo grau de erro, o diagnóstico dos pacientes baseando-se nas mutações (dos mesmos), todas as mutações que não tinham sua localidade cromossômica definida foram excluídas, pois não traziam informação genética para o trabalho. Desta forma, a primeira etapa da seleção, não utiliza nenhum conceito estatístico e o resultado é a exclusão de 2363 mutações, restando ainda 48820 características, conforme Figura 8.

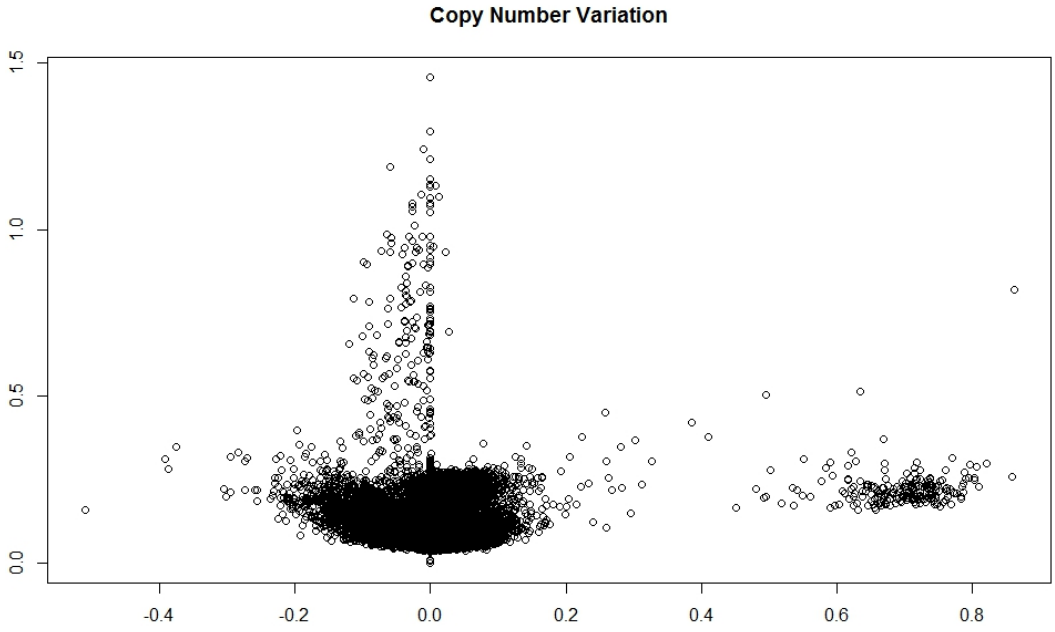

Figura 8. Gráfico que mostra a distribução das características após a primeira etapa de seleção. O eixo X é formado pela mediana e o eixo Y pelo IQR.

A segunda etapa da seleção é focada na variação das mutações entre as amostras. A predição é o resultado da análise das características da amostra por parte do modelo. Se as características tem valores parecidos, então fica difícil para o modelo diferenciar as amostras entre as classes, neste caso diagnóstico.

Existem algumas alternativas para medir o nível de variância das características, por exemplo: media, mediana, desvio padrão, IQR. Neste trabalho foi utilizado o IQR. As 2 mil características com maior IQR permanecem para posteriores seleções e as demais são excluídas, conforme Figura 9.

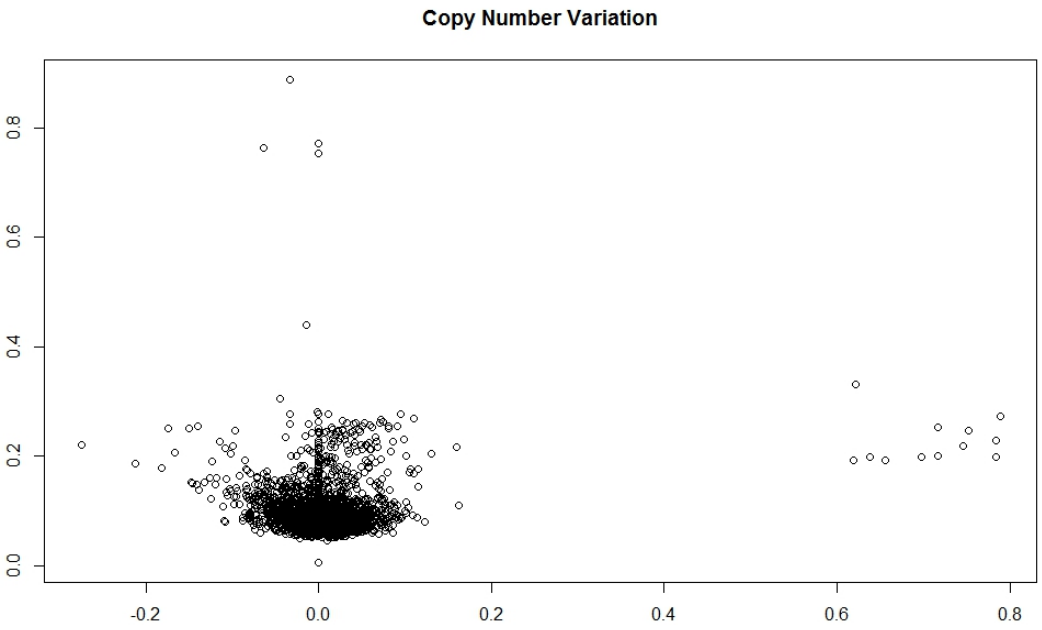

Figura 9. Gráfico que mostra a distribução das características após a segunda etapa de seleção. O eixo X é formado pela mediana e o eixo Y pelo IQR.

A terceira etapa consiste da aplicação da técnica de seleção chamada de ANOVA. O IQR mede a variação genérica das características sem se importar com a variação entre grupos isolados. Por exemplo, dado que uma característica x possa pertencer a 3 classes diferentes A, B e C, x pode ter uma variação alta entre A e B, mas não entre B e C. O filtro ANOVA é utilizado para resolver este problema, priorizando caracaterísticas que variem entre todas as classes. O resultado é a exclusão de 937 características restando ainda 1.063 para a última etapa, conforme Figura 10.

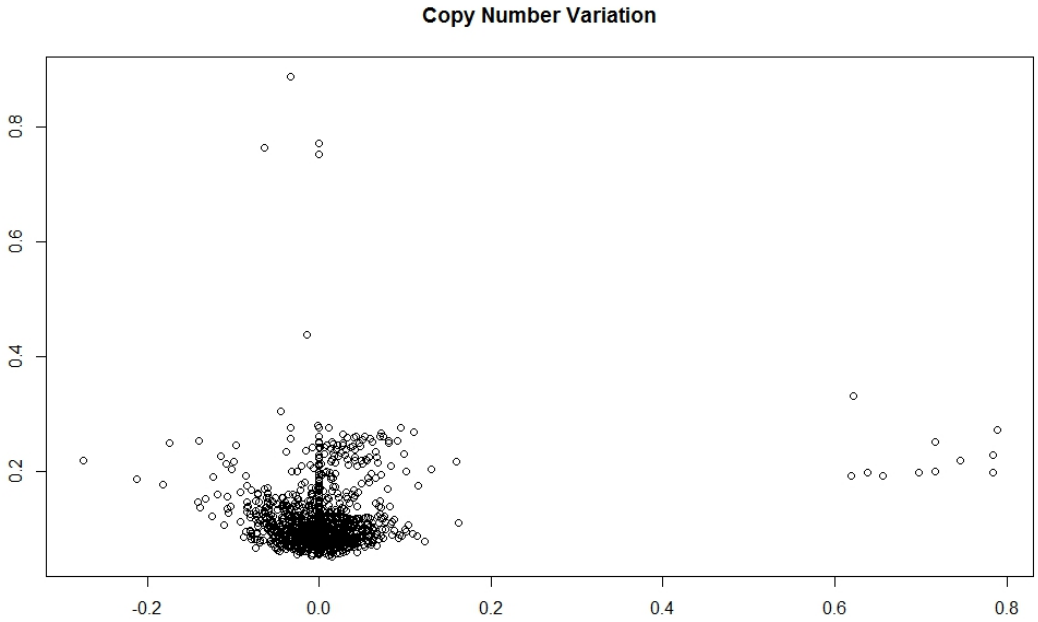

Figura 10. Gráfico que mostra a distribução das características após a terceira etapa de seleção. O eixo X é formado pela mediana e o eixo Y pelo IQR.

O quarto e último filtro utiliza da capacidade da árvore de decisão de medir a importância de uma característica pela perda de entropia, ou ganho de informação, que o sistema sofre quando ela é retirada do sistema.

A aplicação de um modelo chamado de *Random Forest* que é baseado em árvores de decisão, exclui 1033 características, conforme Figura 11. Portanto, é o fim da etapa de seleção.

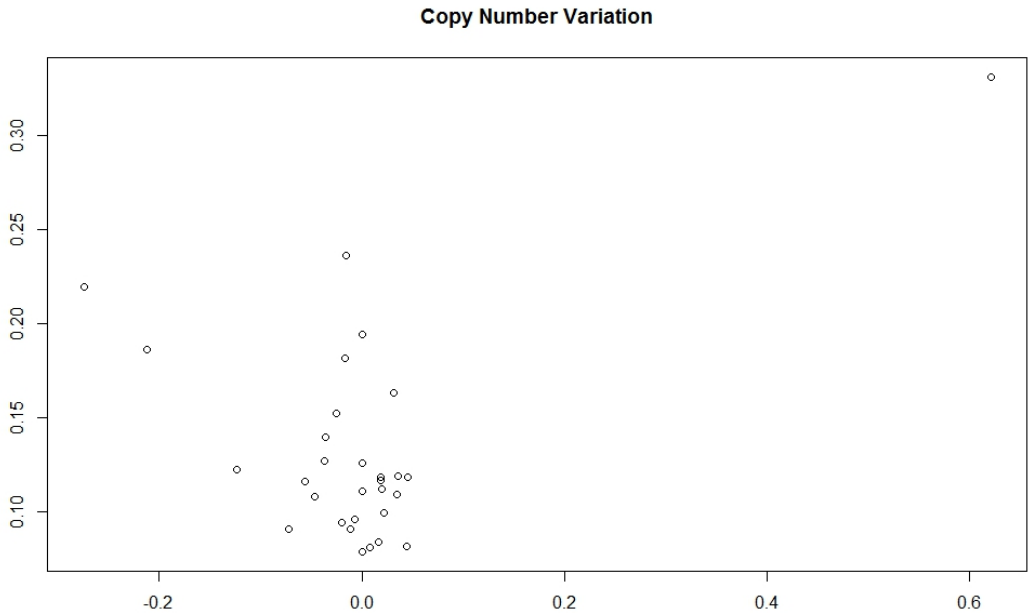

Figura 11. Gráfico que mostra a distribução das características após a última etapa de seleção. O eixo X é formado pela mediana e o eixo Y pelo IQR.

4.3 Validação dos modelos

Os modelos de predição, random forest, KNN e SVM, são configurados utilizando o novo conjunto de dados com 483 amostras e 30 características.

Validação cruzada é então utilizada para avaliar a capacidade de predição dos modelos bem como sua capacidade de generalização. O banco é dividido em 70% de dados para treinamento e 30% de dados para teste. São gerados 30 modelos diferentes, para cada tipo de modelo, com conjuntos de treinamento e de teste diferentes para que, ao final dos testes, possa ser feita uma análise que tenha como entrada a resposta dos modelos para várias configurações de dados diferentes, dessa forma evitando resultados viciados.

Resultados e Discussão

Pode-se observar uma taxa de acerto média superior a 84% para os três modelos, onde os melhores resultados foram encontrados utilizando-se SVM com valor máximo de acerto de 0,944444, média de 0,905556 e mínimo de 0,847222.

O segundo melhor modelo, em média de resultados, foi o *Random Forests* apresentando máximo de acerto de 0,944444, média de 0,887963 e mínimo de 0,847222.

Knn ficou com os resultados menos promissores para este conjunto de dados, apresentando máximo de acerto de 0,8958333, média de 0,8462963 e mínimo de 0,7986111.

A análise completa dos resultados fica mais clara na Figura 12 que mostra um gráfico de bloxpot para cada modelo.

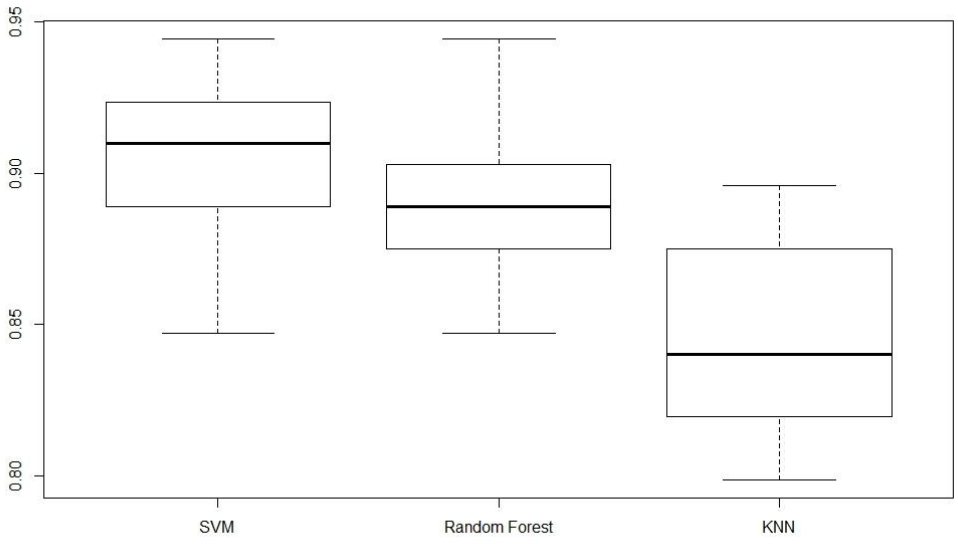

Figura 12. Boxplot que compara os três modelos e seus resultados, SVM, Random Forests e KNN respectivamente.

Para validação estatística dos resultados das predições, foi utilizado o teste não-paramétrico da Soma dos Postos de Wilcoxon com um nível de significância de 99%, conforme Figura 13.

	SVM	Random Forest	KNN
SVM		▲	▲
Random Forest	▽		▲
KNN	▽	▽	

Figura 13. Tabela gerada pelo teste de Wilcoxon.

Os resultados mostram que é possível predizer o diagnóstico de um paciente epilético com erro médio de 10% para o melhor modelo. Atualmente, a forma mais convencional de diagnóstico desta doença acontece por meio de um exame do cérebro chamado eletroencefalograma, que registra atividades de correntes elétricas no encefalo. Existem ainda outros exames e testes como MRI e biópsia cerebral, contudo frequentemente fornecem informações incertas para um diagnóstico e podem ser excessivamente invasivos. Em muitos casos o resultado desses exames é inconclusivo [16].

Este trabalho não tem o intuito de propor, através dos resultados encontrados, um modelo que substitua as técnicas de diagnóstico convencionais, mas sim mostrar que é viável e promissor o estudo de novas tecnologias que diagnostiquem um epilético, e mais futuramente a suscetibilidade à doença, através de variações genéticas sejam elas CNVs, mutações de nucleotídeos, inversões, ou qualquer outra possível.

Conclusões e Trabalhos Futuros

Neste capítulo são feitas contribuições finais sobre a metodologia empregada neste trabalho, seus resultados, suas consequências e possíveis trabalhos futuros.

6.1 Contribuições e Conclusões

A realização do diagnóstico da epilepsia antes que os pacientes sofram a primeira convulsão e ainda, a criação de novas formas de diagnósticos que sejam mais eficiente que os usuais ou que sirvam para complementá-los, é uma contribuição relevante e que pode tornar possível o começo do tratamento antes que a doença comece a trazer danos.

Este trabalho propõe uma abordagem estatística-computacional com modelos que aprendem padrões para que seja possível o diagnóstico de pacientes epiléticos utilizando apenas suas variações genéticas. Para isto, foi utilizado um banco de dados com 483 amostras que contém dados de pacientes com três diagnósticos diferentes: epiléticos com transtorno biolar, epiléticos esquizofrênicos e pacientes em estado controlado.

Os modelos são gerados utilizando 30 das aproximadas 51 mil variações disponibilizadas pelo banco e geram predições com médias de precisão de 0,905556, 0,887963 e 0,8462963, o que demonstra que pode ser viável a utilização de modelos de *machine learning* no diagnóstico de pacientes epiléticos.

6.2 Trabalhos Futuros

Os resultados de trabalhos que utilizam *Data Mining* dependem antes de tudo de sua base de dados. Existem alguns projetos que coletam variações genéticas de epiléticos e que devem disponibilizar os dados em um futuro não muito distante. Desta forma, as análises podem evoluir de diagnósticos para estudo de causa, que foquem não na predição ou classificação das amostras, mas que, através de técnias

de *machine learning*, busquem padrões que possam separar conjuntos de variações genéticas que tenham frequência relevante em epiléticos.

A caracterização da epilepsia genética é um problema complexo e que persiste sem solução há vários anos. Porém, as abordagens para este tipo de pesquisa estão mudando, por exemplo, de análises focadas em um único nucleotídeo para conjuntos de mutações.

Bibliografia

[1] TORGO, L. **Data Mining With R.** Chapman & Hall/CRC, 2011. 277p.

[2] GRIFFTHS, A. **An introduction to Genetic Analysis**. W H Freeman & Company, 1996. 706 p.

[3] FOX, S. **Human Physiology**. McGraw-Hill Science/Engineering/Math, 2010. 832p.

[4] IONITA-LAZA I.; ROGERS, A.; LANGE, C.; RABY, B. e LEE, C. Genetic Association analysis of copy-number variation (CNV) in human disease pathogenesis. **Genomics**, 2008, Boston, EUA

[5] VISSERS, L.; VELTMAN, J.; KESSEL, A. G. e BRUNNER, H. Identification of disease genes by whole genome CGH arrays. **Human Molecular Genetics**, 2005.

[6] PICARD, F.; ROBIN, S.; LAVIELLE, M.; VAISSE, C. e DAUDIN, J. A statistical Approach for array CGH data analysis. **BMC Bioinformatics**, 2005, UK.

[7] DOELKEN, S.; KOHLER, S.; MUNGALL, C.; GKOUTOS, G.; RUEF, B.; SMITH, S.; SMEDLEY, D.; BAUER, S.; KLOPOCKI, E.; SCHOFIELD P.; WESTERFIELD , M.; ROBINSON, P. e LEWIS S. Phenotypic overlap in the contribution of individual genes to CNV pathogenicity revealed by cross-species computational analysis of single-gene mutations in humans, mice and zebrafish. **Disease Models & Mechanisms**. 2013.

[8] FEUK, L.; MARSHALL, C. R.; WINTLE, R. F. e SCHERER S. W. Structural Variants: changing the landscape of chromossomes and design of disease studies. **Human Molecular Genetics**, 2006.

[9] GILLHAM, N. **Genes, Chromosomes and Disease**. Pearson Education, Inc. 2011. 352p.

[10] Feuk, L.; Carson, AR e Scherer, SW. Structural variation in the human genome. **Nature Reviews Genetics**. 2006.

[11] MOHAMMED, S. e OGILVIE, C. Understanding chromossome disorders. **Unique**, 2013, Caterham, Surrey, UK.

[12] OTTMAN, R. Genetic Epidemology of Epilepsy. **Epidemologic Reviews**, 1997, USA

[13] KALACHIKOV, S.; EVGRAFOV, O.; ROSS, B.; WINEWER M.; BARKER-CUMMINGS, C.; BONESCHI, F. M.; CHOI, C.; MOROZOV, P.; DAS, K.; TEPLITSKAYA, E.; YU, A.; CAYANIS, E.; PENCHASZADEH, G.; KOTTMANN, A. H.; PEDLEY, T.; HAUSER, A.; OTTMAN, R. e GILLIAM C. Mutations in LGI1 cause autosomal-dominant partial epilepsy with auditory features. **Nature Genetics**. 2002.

[14] BONAVENTURA, C.; OPERTO, F. F.; BUSOLIN, G.; EGEO, G.; D'ANIELLO, A.; VITELLO, L.; SMANIOTTO, G.; FURLAN, S.; DIANI, E.; MICHELUCCI, R.; GIALLONARDO, A. T.; COPPOLA, G. e NOBILE, C. Low penetrance and effect on protein secretion of LGI1 mutations causing autosomal dominant lateral temporal epilepsy. **International League Against Epilepsy**. 2011.

[15] HWANG, S. e HIROSE S. Genetics of temporal lobe epilepsy. **Brain & Development**. 2012.

[16] HIGGINS, J. J. Autoimmune epilepsy diagnostics: an interview. **Athena Diagnostics**, 2014.

[17] ZAKI, M. J. e MEIRA, W. **Data Mining and Analysis**. Cambridge University Press, 2014. 594p.

[18] NG, A**. Machine Learning, Part V**, Disponível em: http://cs229.stanford.edu/notes/cs229-notes3.ps. Acesso em: 02/12/2014.

[19] MITCHELL, T. **Chapter 3 - Decision Tree Learning**, Disponível em: http://cs229.stanford.edu/notes/cs229-notes3.pdf. Acesso em: 02/12/2014.

[20] RAFAEILZADEH, P.; TANG, L. e LIU, H. Cross-Validation. **Encyclopedia of Database Systems**, 2009, pp 532-538.

[21] KIRKMAN, T. W. Statistcs to use. 1996.

[22] BAUDET, A.; Burant-Hall, A. e STEWART, L. High Frequency of CNV Mutations in Combined Schizophrenia and Epilepsy. Disponível em: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE23703. Acesso em: 02/12/2014.

[23] FAWCETT, T. An Introduction to ROC analysis. Disponível em: https://ccrma.stanford.edu/workshops/mir2009/references/ROCintro.pdf. Acesso em: 21/12/2014.

[24] SIEGERT, S. Variance estimation for Brier Score decomposition. Disponível em: http://arxiv.org/pdf/1303.6182.pdf. Acesso em: 21/12/2014.