CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DE MINAS GERAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM MATEMÁTICA E COMPUTACIONAL

CLASSIFICADORES PARA PACIENTES COM CÂNCER DE MAMA DE ACORDO COM A SENSIBILIDADE QUIMIOTERÁPICA NEOADJUVANTE

PEDRO KÁSSIO RIBEIRO MATOS LOUREIRO DE CARVALHO

BELO HORIZONTE
AGOSTO DE 2017

PEDRO KÁSSIO RIBEIRO MATOS LOUREIRO DE CARVALHO

CLASSIFICADORES PARA PACIENTES COM CÂNCER DE MAMA DE ACORDO COM A SENSIBILIDADE QUIMIOTERÁPICA NEOADJUVANTE

Projeto de Qualificação apresentado ao Programa de Pós-graduação em Modelagem Matemática e Computacional do Centro Federal de Educação Tecnológica de Minas Gerais, como requisito parcial para a obtenção do título de Mestre em Modelagem Matemática e Computacional.

Área de concentração: Modelagem Matemática e Computacional

Linha de pesquisa: Sistemas Inteligentes

Orientador: Thiago de Souza Rodrigues

CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DE MINAS GERAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM MATEMÁTICA E COMPUTACIONAL
BELO HORIZONTE
AGOSTO DE 2017

Carvalho, Pedro Kássio Ribeiro Matos Loureiro de
C331c Classificadores para pacientes com câncer de mama de acordo com a sensibilidade quimioterápica neoadjuvante / Pedro Kássio Ribeiro Matos Loureiro de Carvalho. – 2017.
xi, 45 f. : il.

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Modelagem Matemática e Computacional.
Orientador: Thiago de Souza Rodrigues.
Dissertação (mestrado) – Centro Federal de Educação Tecnológica de Minas Gerais.

1. Quimioterapia – Modelos matemáticos – Teses. 2. Câncer – Tratamento adjuvante – Teses. 3. Bioinformática – Teses. 4. Redes neurais – Teses. 5. Máquina de aprendizagem extrema – Teses. 6. Inteligência coletiva – Teses. I. Rodrigues, Thiago de Souza. II. Centro Federal de Educação Tecnológica de Minas Gerais. III. Título.

CDD 511.8

Elaboração da ficha catalográfica pela Biblioteca-Campus II / CEFET-MG

SERVIÇO PÚBLICO FEDERAL
MINISTÉRIO DA EDUCAÇÃO
CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DE MINAS GERAIS
COORDENAÇÃO DO PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM MATEMÁTICA E COMPUTACIONAL

ATA DA 244ª SESSÃO PÚBLICA DE APRESENTAÇÃO E DEFESA DE DISSERTAÇÃO PARA OBTENÇÃO DO TÍTULO DE MESTRE EM MODELAGEM MATEMÁTICA E COMPUTACIONAL. Em 30 de agosto de 2017, no Auditório 401 do Prédio 17 – Departamento de Computação – DECOM do Centro Federal de Educação Tecnológica de Minas Gerais – CEFET-MG, reuniu-se, às 14 horas, a Banca Examinadora, designada para este fim, pelo Colegiado do Programa de Pós-Graduação em Modelagem Matemática e Computacional, constituída pelos professores: Dr. Thiago de Souza Rodrigues (Orientador), Dr. João Fernando Machry Sarubbi e Dr. Vinícius Fernandes dos Santos, para examinar o trabalho do mestrando **Pedro Kássio Ribeiro Matos Loureiro de Carvalho,** sob o título "***CLASSIFICADORES PARA PACIENTES COM CÂNCER DE MAMA DE ACORDO COM A SENSIBILIDADE QUIMIOTERÁPICA NEOADJUVANTE***". O Prof. Dr. Thiago de Souza Rodrigues, como presidente da Banca Examinadora, declarou aberta a sessão, passando a palavra ao candidato, para que este expusesse sua dissertação. Terminada a exposição, o Presidente passou a palavra aos membros da Banca Examinadora, que iniciaram a arguição, na seguinte ordem: Prof. Dr. João Fernando Machry Sarubbi, Prof. Dr. Vinícius Fernandes dos Santos e Prof. Dr. Thiago de Souza Rodrigues. Terminada a arguição, retirou-se a banca examinadora para deliberação. De volta ao recinto, o Presidente deu conhecimento ao candidato de que sua Dissertação foi aprovada, devendo a redação final incorporar as contribuições da banca examinadora no prazo máximo de 60 (sessenta) dias. Nada mais havendo a tratar, o Presidente declarou encerrada a sessão, cujas atividades são registradas nesta Ata, lavrada, a qual assina, juntamente com os demais membros da Banca Examinadora. Belo Horizonte, 30 de agosto de 2017.

Prof. Dr. Thiago de Souza Rodrigues
Centro Federal de Educação Tecnológica de Minas Gerais

Prof. Dr. João Fernando Machry Sarubbi
Centro Federal de Educação Tecnológica de Minas Gerais

Prof. Dr. Vinícius Fernandes dos Santos
Centro Federal de Educação Tecnológica de Minas Gerais

SERVIÇO PÚBLICO FEDERAL
MINISTÉRIO DA EDUCAÇÃO
CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DE MINAS GERAIS
COORDENAÇÃO DO PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM MATEMÁTICA E COMPUTACIONAL

"CLASSIFICADORES PARA PACIENTES COM CÂNCER DE MAMA DE ACORDO COM A SENSIBILIDADE QUIMIOTERÁPICA NEOADJUVANTE"

Dissertação de Mestrado apresentada por **Pedro Kássio Ribeiro Matos Loureiro de Carvalho**, em 30 de agosto de 2017, ao Programa de Pós-Graduação em Modelagem Matemática e Computacional do CEFET-MG, e aprovada pela banca examinadora constituída pelos professores:

Prof. Dr. Thiago de Souza Rodrigues (Orientador)
Centro Federal de Educação Tecnológica de Minas Gerais

Prof. Dr. João Fernando Machry Sarubbi
Centro Federal de Educação Tecnológica de Minas Gerais

Prof. Dr. Vinícius Fernandes dos Santos
Centro Federal de Educação Tecnológica de Minas Gerais

Visto e permitida a impressão,

Prof. Dr. José Geraldo Peixoto de Faria
Coordenador do Programa de Pós-Graduação em
Modelagem Matemática e Computacional

Agradecimentos

Eu agradeço especialmente ao meu orientador, Thiago de Souza Rodrigues, que me ajudou em todas as etapas desse trabalho. Me acompanhou desde a graduação e agora nessa nova conquista. Sem sua ajuda nada disso seria possível. Agradeço também à CAPES e ao CEFET-MG. O primeiro pela ajuda de custo com bolsa de estudos e o segundo pela infraestrutura que permitiu que o meu trabalho fosse desenvolvido. Agradeço aos meus professores que me ajudaram tirando as dúvidas em relação às disciplinas feitas para que o projeto pudesse ser executado. Agradeço aos professores que participaram da banca avaliadora, por me ajudarem a melhorar o trabalho e por aceitarem o convite de avaliar minha dissertação. Agradeço também aos meus familiares e amigos que me ajudaram em cada etapa, me dando força para continuar mesmo diante das dificuldades. Agradeço especialmente à minha mãe, meu pai, minha irmã e minha namorada, que sempre estiveram próximos e me ajudaram a melhorar meu trabalho e a ter mais força a cada dia. Eles foram vitais para que todo o esforço se torna-se realidade, sempre me apoiaram e me ajudaram e foram o principal pilar para esse objetivo.

Resumo

O câncer de mama é tipo de câncer mais comum entre as mulheres. O seu tratamento é geralmente efetuado por meio da utilização de quimioterapia neoadjuvante, também conhecida como pré-operatória, seguida da operação para remoção do tumor e posteriormente a quimioterapia pós-operatória. A quimioterapia neoadjuvante pode ter Resposta Patológica Completa (PCR), quando a doença é eliminada, ou pode possuir Doença Residual (RD). Este projeto busca utilizar informações de subtipos moleculares do câncer de mama, para identificar cinco tipos específicos destas expressões gênicas e, por meio da implementação de diversos tipos diferentes de classificadores, avaliar o efeito da quimioterapia neoadjuvante nos portadores da doença. O melhor classificador obtido foi o *Extreme Learning Machine*, que possui um tempo de execução baixo e 80% de acerto em média, indicando um bom resultado.

Palavras-chave: Neoadjuvante, PCR, RD, Câncer, Bioinformática, Redes Neurais, Extreme Learning Machine, Particle Swarm Optimization.

Abstract

The breast cancer is the most common type of cancer one on women. This cancer treatment is usually done with neoadjuvant chemotherapy, also known as pre-operation chemothrerapy, followed by the operation to remove the tumor and aftar the post operation chemotherapy. The neoadjuvant chemotherapy may show Complete Pathological Response (CPR), when the disease is eliminated, or Residual Disease (RD). This project seeks to use information of the molecular subtypes of breast cancer, to identify five specific types of those genic expressions and by implementing several different types of classifiers evaluate the effect of neoadjuvant chemotherapy in patients with the disease. The best classifier was obtained using Extreme Machine Learning algorithm, which has a very small runtime and 80% of accuracy on average, indicating a good result.

Keywords: Neoadjuvant, CPR, RD, Cancer, Bioinformatics, Neural Networks, Extreme Learning Machine, Particle Swarm Optimization.

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

PCR	Resposta Patológica Completa
RD	Doença Residual
RNA	Redes Neurais Artificiais
MLP	Multi-Layer Perceptron
PSO	Particle Swarm Optimization
ELM	Extreme Learning Machine

Sumário

1 Introdução

Câncer de Mama é o câncer mais comum em mulheres, sua incidência tem crescido nos países subdesenvolvidos devido ao aumento da expectativa de vida, aumento da urbanização e adoção do estilo de vida ocidental (tipo de alimentação, rotina de trabalho e condições ambientais em geral) (INCA, 2016). Em 2011 foi estimado que mais de 508.000 mulheres morreram devido a esta enfermidade. Embora seja considerada uma doença do mundo desenvolvido, quase 50% dos casos e 58% das mortes ocorrem em países menos desenvolvidos. Diversos estudos demonstram que o câncer de mama foi a principal causa de morte, em mulheres de 30 a 49 anos, por neoplasia entre 1991 a 1995.

O câncer de mama, segundo o Instituto Nacional de Câncer (INCA), é o tipo de câncer mais comum entre mulheres no Brasil e no mundo, correspondendo a cerca de 25% dos novos casos a cada ano. Em homens ele representa 1% dos casos de câncer, uma ocorrência bem menor, porém existente, sendo considerado relativamente raro antes dos 35 anos e sua ocorrência cresce com a idade, principalmente após os 50 anos. A estimativa é de que o número de novos casos em 2016 foi de aproximadamente 57.960, com 14.388 mortes, das quais 181 são homens e 14.206 mulheres (INCA, 2016).

Devido a elevada incidência de câncer de mama em mulheres, existem diversos estudos referentes a sua prevenção e tratamento. No âmbito de prevenção existem campanhas elaboradas pelo governo que promovem o autoexame e mamografia após os 40 anos, o que tem elevado a taxa de diagnóstico precoce, muito importante para que o tratamento tenha melhor eficácia. No tratamento do câncer de mama, assim como os demais, é ideal que o carcinoma (tumor maligno epitelial ou glandular, que tende a invadir tecidos circundantes, originando metástases) seja detectado nos estágios iniciais da doença, para que as chances de cura sejam maiores e o tratamento irá depender de vários fatores patológicos do crescimento do câncer (neoplasia) (ELSTON; ELLIS, 1991). O tratamento é dividido em três etapas básicas. A primeira corresponde à quimioterapia pré-operatória, também conhecida como neoadjuvante. A segunda etapa consiste no procedimento cirúrgico para remover o restante do tumor que permaneceu após a quimioterapia neoadjuvante. E por fim, a terceira etapa é a quimioterapia pós-operatória, ou adjuvante, que deverá remover qualquer célula cancerígena que tenha permanecido no organismo (SOCIETY, 2017).

1.1 Sensibilidade à Quimioterapia Neoadjuvante

O tratamento do câncer de mama consiste resumidamente em quimioterapia pré-operatória (neoadjuvante), cirurgia e quimioterapia pós-operatória (adjuvante). Em casos operáveis de câncer, o objetivo da quimioterapia neoadjuvante é a redução do tamanho

do tumor a fim de facilitar os próximos estágios do tratamento. A maior parte dos ensaios clínicos que examinaram a correlação entre a resposta patológica completa e a sobrevida livre de câncer a longo prazo têm relatado uma forte associação entre estes dois resultados (GRALOW et al., 2008; MCVEIGH et al., 2014; JACOBS; SIMOS; CLEMONS, 2014; SOTIRIOU; PUSZTAI, 2009). A Resposta Patológica Completa (*PCR - Pathological Complete Response*), é associada aos pacientes que não exibem nenhum tumor residual após a quimioterapia neoadjuvante, apresentando uma excelente sobrevida livre de câncer, ao passo que a Doença Residual (*RD - Residual Disease*) é associada com paciente possuindo qualquer tumor residual (FISHER et al., 1998; KUERER et al., 1999). Devido a possibilidade de erradicar a doença por completo, a quimioterapia neodjuvante é frequentemente prescrita em muitos casos, entretanto a resposta patológica completa ocorre apenas na minoria dos pacientes. O entendimento dos principais fatores que afetam a sensibilidade à quimioterapia (quimiosensitividade) por meio de estudos dos genes expressos é uma importante questão em Oncologia e na Bioinformática Médica, o que torna necessário a construção de um preditor confiável para identificar pacientes PCR e RD, fornecendo um tratamento individual e efetivo.

Os preditores são utilizados para estimar o comportamento de uma resposta baseado nos dados utilizados como entrada do sistema. São muito comuns em algoritmos de inteligência computacional e aprendizado de máquina. Recentemente vêm sendo utilizados na medicina moderna como estratégia para prever o comportamento gênico de células cancerosas.

O estudo dos genes envolvidos nas respostas ao tratamento neoadjuvante de neoplasias é de extrema importância para a classificação dos pacientes. Os genes são classificados cientificamente como sequências de DNA cromossômico que são necessárias para a produção de produtos funcionais, sejam eles polipeptídeos ou moléculas funcionais de RNA. Cada gene codifica um polipeptídeo diferente que irá influenciar no fenótipo do indivíduo. Essa codificação também é chamada de expressão gênica, ou seja, quando um gene expressa moléculas ou polipeptídeos que irão influenciar nas características físicas daquele indivíduo (SNUSTAD; SIMMONS, 2017).

O fenótipo de neoplasia é definido pelo conjunto de mutações e alterações epigenéticas que podem afetar modelos transcricionais e funções proteicas. A transcrição, que é a primeira etapa da expressão gênica, consiste na produção de uma molécula de RNA mensageiro que será utilizada posteriormente na tradução, onde esse RNA será traduzido em polipeptídeos, e consequentemente, irá alterar o fenótipo do indivíduo. Quando os modelos transcricionais são afetados, seja por uma mutação (quando ocorre troca ou deleção de nucleotídeos no DNA), ou por alterações epigenéticas (modificações no genoma que não envolvem mudanças na sequência do DNA) uma expressão gênica exacerbada pode ocorrer, levando ao surgimento de neoplasias. Do mesmo modo, alterações nas funções

proteicas das células podem conferir vantagens competitivas, podendo também levar a célula a expressar um fenótipo maligno (COLOMBO; RAHAL, 2010).

Os subtipos de câncer selecionados para serem utilizados no presente trabalho, foram classificados de acordo com suas expressões gênicas e agrupados em *clusters* por meio das similaridades encontradas em cada gene relacionado ao câncer de mama utilizando a técnica de *microarray*.

A tecnologia de *microarray*, uma ferramenta de análise de expressões gênicas, permite investigar a expressão de centenas ou milhares de genes em uma amostra. O estudo simultâneo de centenas de genes transformou a técnica de *microarray* em uma ferramenta de análise global muito importante, com aplicações em várias áreas, incluindo o estudo do desenvolvimento de neoplasias (COLOMBO; RAHAL, 2010).

A utilização de decisores clínico-patológicos (tais como idade, estágio da doença, sexo, resposta imunológico e dentre outros) fornece informação insuficiente para prever pacientes PCR e RD, entretanto a utilização das medidas de expressões gênicas tem tornado possível a especificação de um perfil genético entre os pacientes, conduzindo a análise de quimiosensitividade para a análise de dados genômicos e dessa forma tornar viável a classificação das possíveis respostas de cada paciente à terapia neoadjuvante, antes mesmo de submetê-lo ao tratamento (MODLICH et al., 2005; GONG et al., 2007; BUCHHOLZ et al., 2002; HESS et al., 2006; BENSMAIL; HAOUDI, 2003; PUSZTAI et al., 2004; PAIK et al., 2004; PUSZTAI et al., 2006; ANDRE et al., 2006).

O tratamento neoadjuvante de câncer de mama geralmente é feito por meio da combinação de dois ou mais fármacos, o que acarreta diversos efeitos colaterais, porém sua eficácia é maior quando administrados dessa maneira. Os principais fármacos utilizados na quimioterapia neodjuvante são:

1. Antraciclinas (exemplos:doxorrubicina e epirrubicina);
2. Taxanos (exemplos: como paclitaxel e docetaxel);
3. 5-fluorouracilo;
4. Ciclofosfamida;
5. Carboplatina

Os efeitos colaterais que podem ser acarretados pela sua utilização são: perda de cabelo, alterações nas unhas, feridas na boca, perda ou aumento do apetite, náuseas e vômitos, diarreia, infecções ()ocasionadas pela diminuição dos glóbulos brancos), hematomas ou hemorragias (ocasionados pela diminuição das plaquetas), fadiga (acarretada pela diminuição dos glóbulos vermelhos) (SOCIETY, 2017).

Atualmente, o câncer de mama é reconhecido como uma doença complexa e clinicamente heterogênea. Ela não é totalmente correlacionada com a evolução clínica, baseada nas classificações histológicas, de acordo com a velocidade de multiplicação

celular (PUSZTAI et al., 2006; JEMAL et al., 2011; CAVALLARO et al., 2012; DONAHUE; GENETOS, 2013). A complexidade do câncer de mama pode ser percebida através dos perfis de expressão gênica por *microarray*, e podem ser utilizados para fornecer informações de prognóstico e também de avaliação clínica padrão (SØRLIE et al., 2001; LOI et al., 2007; PARKER et al., 2009).

O principal objetivo deste trabalho foi criar um classificador para a primeira etapa do tratamento, de forma a identificar se o paciente irá apresentar Resposta Patológica Completa (PCR), ou seja, eliminação total da doença, ou se ele irá apresentar Doença Residual (RD), que ocorre quando o tumor não é eliminado totalmente. No segundo caso não é interessante a aplicação do tratamento neoadjuvante, pois na maioria dos casos o tratamento não irá apresentar efeitos relevantes e poderá causar efeitos colaterais que poderão deixar o indivíduo debilitado para as próximas etapas do tratamento.

Para a seleção de indivíduos PCR ou RD foram construídos classificadores, que são estruturas computacionais que identificam se um indivíduo irá pertencer a uma classe ou outra, de acordo com os dados utilizados de entrada. Neste caso, os dados são as informações de expressões gênicas de cada indivíduo.

Diferenças no número de cópias de genes específicos foram reveladas através de genômica comparativa baseada em dados de *microarray* em diferentes subtipos de cânceres de mama, o que confirma não ser uma doença única com características morfológicas variáveis mas, em vez disso, um grupo de distúrbios neoplásicos distintos molecularmente (CHIN et al., 2006; MELCHOR et al., 2008; CHIN et al., 2007; MELCHOR et al., 2007; BERGAMASCHI et al., 2006; VINCENT-SALOMON et al., 2007; RICHARDSON et al., 2006; SOTIRIOU; PUSZTAI, 2009).

O perfil de expressão gênica tem sido utilizado para distinguir cinco classes moleculares principais de câncer de mama: *Basal-like*, *Luminal-A*, *Luminal-B*, HER2, Normal-like.

O subtipo *Basal-like* é caracterizado pela expressão de vários genes em células basais/mioepiteliais e demonstra padrão prognóstico mais reservado, significando que as chances de sobrevida livre da doença são extremamente pequenas (CIRQUEIRA et al., 2011).

Os subtipos luminais têm classificação proveniente da similaridade que as células neoplásicas desses grupos possuem com as células mamárias normais. O subtipo molecular *Luminal A*, correspondente a cerca de 60% das ocorrências de câncer de mama, apresenta, em relação aos demais, o melhor prognóstico. São tumores, que em sua maioria, apresentam receptor de estrogênio positivo. O aumento dos níveis de estrogênio no organismo pode acarretar o desenvolvimento desse subtipo de câncer. O subtipo Luminal-B é caracterizado pela expressão de genes associados a uma alta proliferação celular, o está associado a um pior prognóstico em relação aos tumores do subtipo Luminal-A (CIRQUEIRA et al., 2011).

O subtipo HER2 apresenta uma expressão elevada da proteína HER2, que aparenta estar associada à resistência à terapias endócrinas. Esse subgrupo apresenta o segundo pior prognóstico quando comparado aos demais (CIRQUEIRA et al., 2011).

O subtipo *Nomal-like* apresenta uma expressão elevada de genes comuns a células epiteliais normais e devido a isso esse subgrupo acaba não apresentando marcadores tumorais comuns (CIRQUEIRA et al., 2011).

Esses cinco subtipos são agrupados em dois grupos, que foram utilizados como entradas de dados para o presente trabalho. Esses grupos são o Basal, composto pelo subtipo *Basal-like*, e o grupo Não-basal, composto pelos subtipos *Luminal-A*, *Luminal-B*, HER2. A partir deste ponto iremos tratar os subtipos baseados nesses grupos.

Ao longo dos anos vários trabalhos de seleção gênica e clusterização foram realizados para chegar aos genes atualmente utilizados para a estratificação dos subtipos. Alguns trabalhos são descritos adiante, demonstrando como se chegou aos subtipos utilizados atualmente.

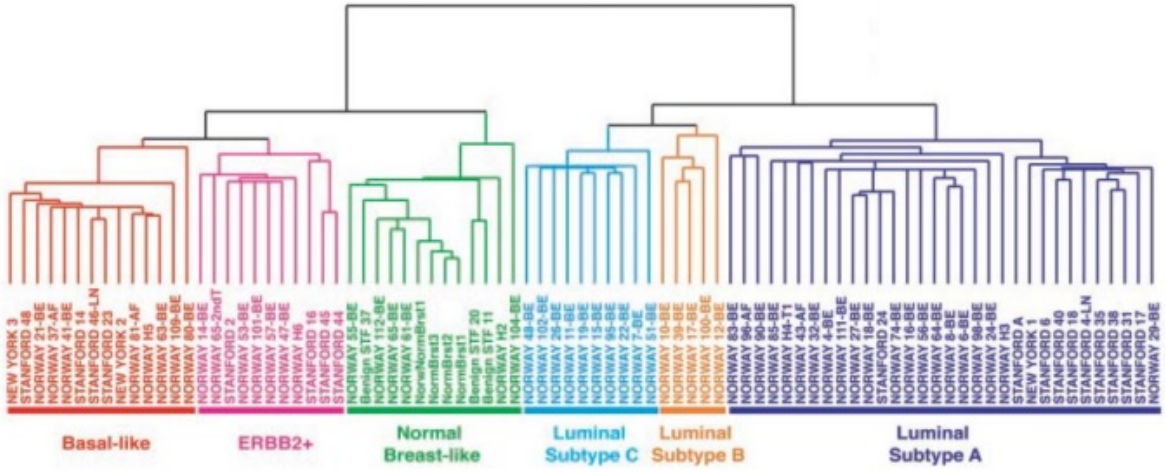
Perou et al. (2000) obteve dados de expressão gênica de 84 amostras, correspondendo à 42 indivíduos. Foram selecionados 8102 genes, dos quais 1753 inicialmente foram considerados. Este conjunto foi então reduzido para 534 genes, que foram denominados subconjunto intrínseco.

Em Sørlie et al. (2001) é mostrado que o grupo *Luminal-A*, que anteriormente representada todo o conjunto *Luminal*, poderia ser dividido em, no mínimo, dois subgrupos com padrões de expressão gênica distintos, e foi separado em Luminal-A, Luminal-B e Luminal-C. A diferenciação entre Luminal-B e Luminal-C é muito pequena e por isso esses subtipos são normalmente agrupados apenas em Luminal-B, como é o caso do presente trabalho. É possível ver a análise feita por clustering hierárquico, das 85 amostras, na Figura 1, onde foram separadas em dois grupos principais. O ramo da esquerda contém três subgrupos, sendo caracterizados pela baixa ou ausência de expressão gênica referente à ER (receptor de estrogênio). E o grupo Luminal que foi separado em três subgrupos distintos (ramo da direita), na Figura 1.

Outra análise por *clustering* foi realizada utilizando o subconjunto de genes intrínseco em um conjunto de 58 amostras, resultando em um dendograma similar ao citado anteriormente. A única diferença foi a posição das cinco amostras Luminal-B, que ficaram agrupadas ao lado do subtipo HER2 (SØRLIE et al., 2001).

Em Parker et al. (2009) a utilidade dos subtipos moleculares de câncer de mama foi estudada individualmente e como parte de um preditor de risco em duas populações: pacientes não recebendo nenhuma terapia sistêmica e pacientes recebendo paclitaxel, fluorouracil, doxorubicin, e cyclophosphamide (T/FAC) como quimioterapia neoadjuvante. Os autores expandiram o subconjunto de genes intrínseco com os genes encontrados em

Figura 1 – As amostras foram divididas em cinco (ou seis) subtipos baseados nas diferenças na expressão gênica.



Fonte: Sørlie et al. (2001)

quatro estudos prévios de microarrays: Sørlie et al. (2001), Hu et al. (2006), Perreard et al. (2006), Sørlie et al. (2003). Quatro grupos adicionais foram identificados como clusters significantemente reproduzíveis. Os quatro subgrupos possuem perfis de expressão gênica similares aos grupos Luminal-A e Luminal-B e poderiam representar estados intermediários, sugerindo novos subgrupos.

Através da utilização do test-t, para cada grupo, foi obtido um conjunto de genes reduzidos. Os subgrupos do câncer de mama foram reproduzidos utilizando os 50 genes através do método *Prediction Analysis of Microarray* (PAM) Tibshirani et al. (2002). Este conjunto foi chamado de PAM50 devido a sua reprodutibilidade na classificação de subtipos.

Em Herschkowitz et al. (2007), os autores reproduziram a classificação dos subtipos moleculares do câncer de mama utilizando o PAM50 em conjunto com dois bancos de dados distintos, um com 240 amostras e outro com 1340 (HOADLEY et al., 2007; KESSLER et al., 2012). Esta reprodutibilidade também foi mostrada em Prat, Ellis e Perou (2012) e Nielsen et al. (2010), e mostra a habilidade do conjunto PAM50 de estratificar os cinco subtipos moleculares do câncer de mama.

A partir desses resultados, os cinco subtipos moleculares do câncer de mama foram utilizados na execução do presente trabalho, separados nos dois agrupamentos citados previamente, para determinar o conjunto de genes adequado para estratificar todos os subtipos com dados de *microarray*, sendo este um dos objetivos. Desta forma será possível criar preditores baseados nesses dados, para que se possa tentar classificar os pacientes como PCR (Resposta Patológica Completa) e RD (doença residual).

Os dados foram separados em dois agrupamentos: basal e não basal. A separação

foi feita de acordo com as suas distribuições e características diferenciadas, conforme citado anteriormente. Além disso, o agrupamento não basal apresenta distribuição binomial, enquanto o basal apresenta distribuição quasi-binomial para este caso, mas na maioria das vezes não apresenta distribuição identificável. Esses fatos levam à utilização dessa divisão em grupos, pois os subtipos dentro de cada grupo são similares entre si. O tipo de distribuição dos agrupamentos será abordado mais adiante.

A seleção das expressões gênicas é fortemente discutida em diversos trabalhos, pois é muito importante que as melhores sondas de DNA utilizadas no *microarray* para classificar os pacientes sejam escolhidas, melhorando a resposta dos classificadores. As sondas são os fragmentos gênicos conhecidos que correspondem aos genes de interesse. Dessa forma é possível identificar a expressão desses genes conhecidos e relacionados a neoplasias nas amostras biológicas (COLOMBO; RAHAL, 2010).

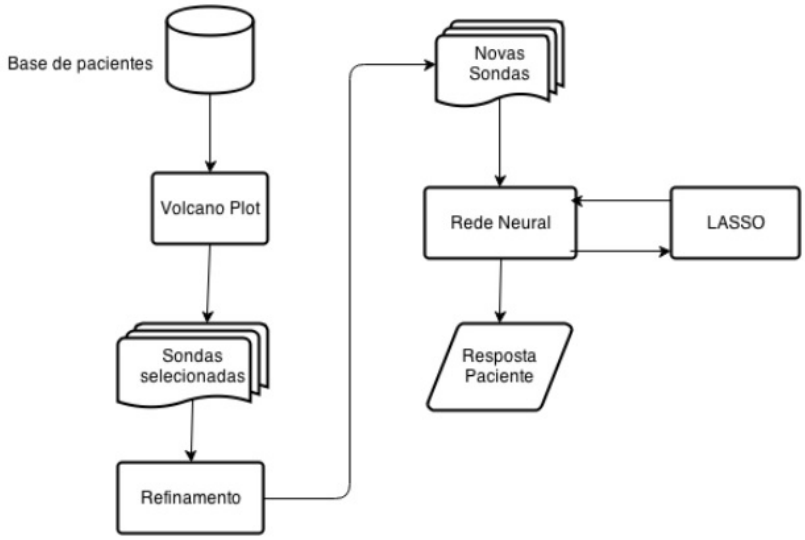
Em Modlich et al. (2005) várias técnicas são aplicadas, baseadas em modelos estatísticos, onde cerca de 59 genes são selecionados com base no *t-test*, e divididas em 31 genes que possuem assinatura mais favorável e 26 menos favoráveis. Os autores conseguiram um resultado de 74% de especificidade (indica a capacidade em identificar corretamente indivíduos que não apresentam a condição de interesse) para o grupo PCR e 62% para RD. Estes valores são razoáveis, porém ainda precisam ser melhorados. Em geral um bom classificador deve possuir uma especificidade de aproximadamente 80% ou mais.

Em Hess et al. (2006) é proposta a construção de um preditor para PCR e RD, comparando diversas técnicas e utilizando uma base de dados de 133 pacientes, que também foi utilizada inicialmente neste trabalho. Posteriormente foi utilizada uma base mais completa desta mesma base, contendo um número maior de pacientes, 345 no total. As principais técnicas desse trabalho são os algoritmos DLDA e vizinho mais próximo. Os melhores resultados obtidos nesse trabalho foram do preditor baseado no DLDA, com sensibilidade de 0,92, especificidade de 0,71 e acurácia de 0,76. Esse classificador faz parte da família de classificadores Nauve Bayes e utiliza a análise diagonal linear discriminante para fazer a separação dos grupos. Nesse classificador a seleção das expressões gênicas foi feita utilizando o *t-test*, com valores de corte baseado no *p-value*, obtendo assim 31 expressões gênicas para utilização no classificador.

Em Horta (2008) o autor trabalhou na redução da quantidade de expressões gênicas, e obteve um resultado semelhante ao anterior. A técnica utilizada para a seleção foi o *Volcano Plot*. O classificador foi o *naive Bayes*, seguida da validação cruzada dos paciente e análise com base na curva ROC. Dois grupos de sondas foram estabelecidos nesse trabalho, um com 18 e outro com 11 sondas. Os resultados obtidos foram semelhantes, porém com uma quantidade menor de sondas. O conjunto de 11 expressões gênicas é um subconjunto das 18 sondas, que foram utilizadas na técnica baseada no algoritmo LASSO.

A metodologia pode ser vista na Figura 2. A base de dados utilizada foi a mesma do trabalho de Hess et al. (2006).

Figura 2 – Metodologia utilizada por Horta (2008)



Fonte: O. (2015)

Baseado nesses trabalhos o objetivo dessa dissertação é utilizar novos métodos para a seleção de características e outros algoritmos conhecidos, porém ainda não utilizados na literatura para tal finalidade, correspondente a implementação do preditor, para buscar resultados melhores na classificação dos pacientes. Novas bases de dados também devem ser levantadas e verificadas, na tentativa de obter uma maior quantidade de dados para os testes e análises. E assim validar a qualidade dos classificadores implementados no presente trabalho. Para aumentar essas bases, será necessário fazer um levantamento de estudos de câncer de mama com o mesmo foco, que gerariam dados semelhantes e associar com a base utilizada atualmente.

A hipótese deste trabalho é que, ao separar os pacientes pelo subtipo molecular, e seus agrupamentos, a classificação desses indivíduos, de acordo com a sensibilidade quimioterápica, será mais eficiente pois cada subtipo possui assinaturas e características gênicas distintas. Entretanto, a estratificação dos pacientes pelo subtipo molecular não é uma tarefa simples, onde conjuntos de genes distintos são apresentados não sendo avaliada a coerência entre esses genes. Utilizando métodos de seleção mais eficientes e preditores mais robustos será possível realizar uma classificação com uma maior sensibilidade e especificidade para resposta patológica à quimioterapia neoadjuvante.

A identificação da sensibilidade à quimioterapia neoadjuvante de um paciente, pode ser utilizada para evitar que indivíduos com uma sem resposta relevante à esse tipo de quimioterapia passem por um processo de grande desgaste, evitando a primeira etapa do

tratamento nesses casos. Assim propomos aplicar novos modelos de classificadores e de seleção de expressões gênicas chaves para a solução do problema.

Esse trabalho consiste na seleção das expressões gênicas mais relevantes para o câncer de mama, baseado em técnicas estatísticas uni e multivariadas, e a partir delas aplicar alguns classificadores. Foram verificadas algumas bases de dados existentes na literatura, para a seleção da mais completa. A partir disso os classificadores foram analisados para verificar se sua resposta é boa ou não para a classificação de paciente do tipo PCR e RD.

O classificador irá funcionar em duas etapas, uma de seleção das melhores expressões gênicas, baseado em métodos de seleção mono variado e multivariado e posteriormente a utilização de algoritmos (estes algoritmos são informados com detalhes na seção de metodologia) com treinamento supervisionado para a classificação dos pacientes.

1.2 Objetivos - Geral e Específicos

1.2.1 Objetivo Geral

Uma das etapas do tratamento do câncer de mama consiste da quimioterapia neoadjuvante que, fornecendo a oportunidade de erradicação da doença, é frequentemente prescrita para muitos pacientes. Como a minoria dos indivíduos apresenta uma resposta completa ao tratamento, o entendimento dos principais fatores que afetam a quimiosensibilidade é uma importante questão em Oncologia e na Bioinformática Médica. Deste modo, nosso objetivo principal é desenvolver um preditor para Sensibilidade Quimioterápica Neoadjuvante do Câncer de Mama, baseado na expressão gênica dos subtipos moleculares de cada agrupamento (basal e não basal), utilizando dados de *microarray*. Além disso, buscamos corroborar com a premissa de que a classificação dos subtipos não basais será melhor em relação aos subtipos basais, pois conforme a literatura eles possuem uma distribuição bem aleatória, o que dificulta na aplicação de classificadores.

1.2.2 Objetivos Específicos

- Determinar o conjunto de genes capazes de estratificar pacientes com Resposta Patológica Completa(PCR) ou Doença Residual(RD) para cada agrupamento dos subtipos moleculares, previamente definidos. Alguns autores, mostraram a utilização de métodos de seleção gênica combinados com classificadores para a predição de indivíduos de acordo com a sensibilidade quimioterápica neoadjuvante. Para isso separamos os pacientes do agrupamento não basal (*Luminal-A*, *Luminal-B*, *HER-2*, *Nomal-like*), dos pacientes do agrupamento basal.

- Aplicar métodos de seleção univariado e multivariado em cada base para seleção dos genes.
- Aplicar métodos de classificação para classificar os pacientes em PCR e RD.

1.3 Organização do Texto

O presente trabalho está dividido em seis capítulos. O atual fornece uma introdução ao tema, com a contextualização, motivação do problema e a descrição dos objetivos proposto para o tratamento do problema em questão.

O segundo capítulo apresenta uma revisão bibliográfica de alguns trabalhos de interesse para a área. Trabalhos que analisaram expressões gênicas e outros que estudaram dados clínicos.

O Capítulo 3 descreve aspectos relevantes para o entendimento, como alguns métodos de seleção de genes e classificação de pacientes, sendo este o de fundamentação teórica para a busca da solução do problema.

O Capítulo 4 demostra os experimentos realizados, juntamente com a análise e discussão dos resultados obtidos.

O Capítulo 5 apresenta a Conclusão do trabalho.

O Capítulo 6 lista os próximos passos do trabalho, apresentando os trabalhos futuros.

2 Fundamentação Teórica

O presente capítulo apresenta alguns fundamentos teóricos necessários para entendimento do restante do trabalho. É explicado como os dados de *microarrays* são obtidos. São explicados de forma resumida alguns algoritmos utilizados para a seleção das expressões gênicas utilizadas como entrada nos algoritmos de classificação, e são explicados os classificadores implementados.

2.1 *Microarrays*

Os dados utilizados no presente trabalho são os de expressões gênicas, coletados com a técnica de *microarrays*. Esse tipo de dado fornece uma medição para os níveis de expressão gênica do indivíduo e pode ser utilizado para criar preditores da resposta patológica do paciente.

Os dados de *microarray* são uma representação numérica para cada expressão gênica do indivíduo, e são obtidos seguindo os seguinte passos:

1. Escolha dos tecidos a serem estudados;
2. Extração de RNA;
3. Preparação da amostra e rotulamento dos genes procurados;
4. Hibridização;
5. Lavagem;
6. Aquisição da imagem.

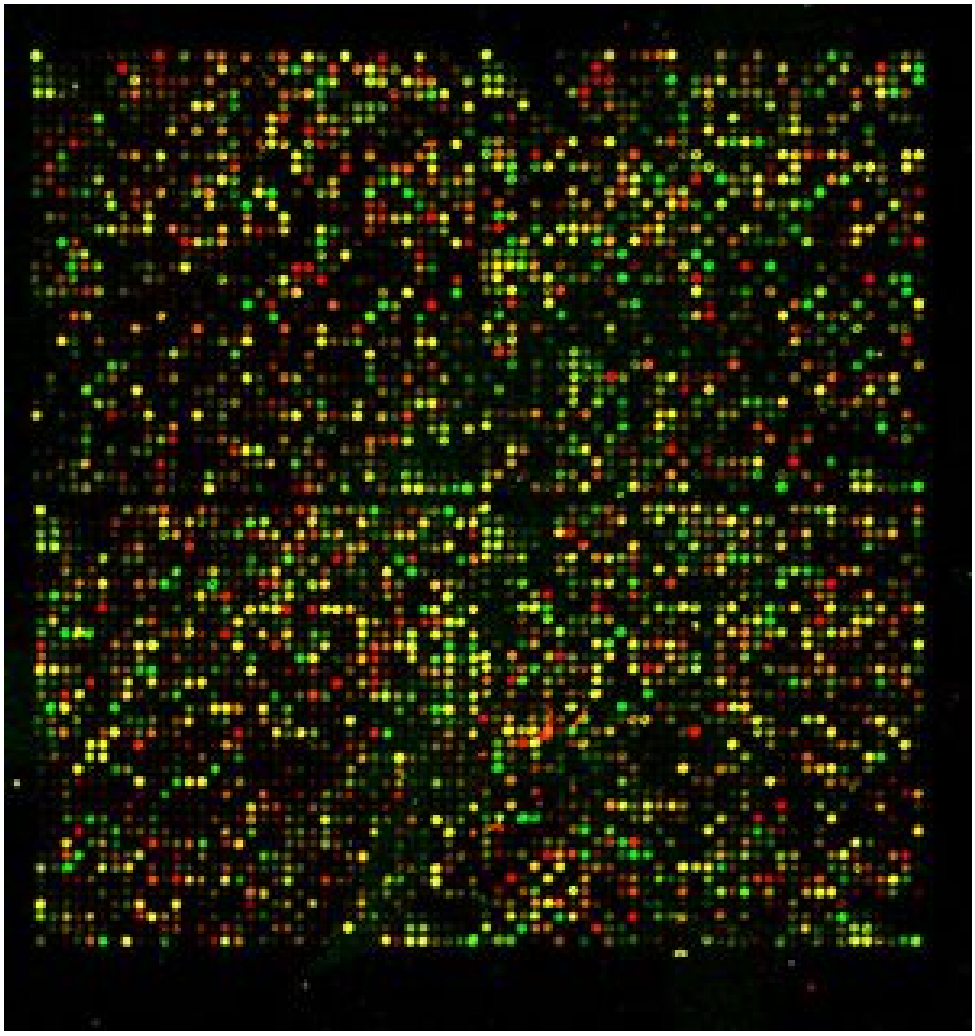
Após a aquisição da imagem, é feita uma leitura da variação de cores presentes na mesma. De acordo com a intensidade da coloração um valor numérico é estabelecido para o nível da expressão daquele gene. Na **??** é apresentado um exemplo de *microarray*, no qual a variação da coloração pode ser claramente observada.

A quantidade de etapas existentes podem gerar um ruído nos dados, algo que deve ser levado em consideração.

Os dados utilizados no presente trabalho são fornecidos por bancos de dados abertos sobre o assunto, e foram obtidos utilizando *biochips* da empresa Affymetrix. Esse tipo de aquisição de dados se baseia na hibridização, seguindo os seguintes passos:

1. Obter um trecho de DNA a partir dos dados coletados pelo *biochip*;
2. Comparar o trecho obtido com o genoma humano para verificar se existe alguma repetição desse trecho em outro gene.
3. Inserir uma molécula de RNA para que o gene seja expresso após o encontro com a

Figura 3 – Exemplo de *microarray*. Imagem obtida em https://www.pinterest.es/explore/dna-microarray/ Acesso em: 10 de setembro de 2017



sonda.

Os dados utilizados foram os presentes em Itoh et al. (2014) e Hatzis et al. (2011), contendo 305 pacientes e 22283 expressões gênicas para cada um deles. Esses pacientes correspondem aos que fazem parte dos subtipos não basais, pois sua configuração de expressões gênicas são mais estáveis e mais conhecidas para a criação dos classificadores.

Os resultados mais recentes, mesmo sendo de aproximadamente 76% de acerto, não foram tão elevados para os classificadores e com isso uma possibilidade para melhoria levantada para o presente trabalho foi a utilização de técnicas de seleção diferenciadas das padrões para esse tipo de problemas, sendo a principal delas o *t-test* e o *Volcano Plot*. Para diferenciar, aplicamos o método *stepwise regression* e o *stepwise GLM*, sendo o último o principal para os resultados discutidos mais adiantes, pois ele é multivariado. No presente trabalho utilizamos todos esses métodos de seleção, sendo o *t-test* utilizado em conjunto com o *Volcano Plot*, para encontrar os *p-values* utilizados no mesmo.

Nas próximas seções serão descritos alguns métodos para seleção das expressões

gênicas, ou sondas, que foram utilizadas no presente trabalho. Este é um dos maiores problemas no presente trabalho, pois as expressões gênicas possuem uma grande influência sobre a resposta do classificador e com isso devem ser muito bem selecionadas. Um grande problema também é a quantidade de sondas utilizadas. Conforme veremos mais adiante, a quantidade de expressões gênicas melhora o resultado do classificador, porém existe um certo limite para que o processamento ocorra sem problemas. Com isso quanto menor a quantidade de expressões, menor o tempo de processamento. Entretanto deve existir uma garantia de qualidade. Isso gera a necessidade de muito cuidado na seleção dos genes.

2.2 Técnicas de Seleção

Esta seção irá apresentar as principais técnicas utilizadas neste projeto para a seleção de expressões gênicas a serem utilizadas nos classificadores.

2.2.1 *T-test*

O *t-test* é uma técnica estatística que compara duas amostras independentes baseadas em suas médias. Esse teste é baseado na Equação 1.

$$y = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_x}{n} + \frac{s_y}{m}}} \tag{1}$$

Nessa equação clássica da estatística, temos o seguinte:

- $\overline{x}$ e $\overline{y}$ são as médias das amostras.
- s_x e s_y são os desvios das amostras.
- n e m são os tamanhos das amostras

Feito o cálculo, é analisado o *p-value*, valor otido para o teste de hipótese baseado em um nível de confiança de 95%, sendo este o limiar de corte para selecionar as características mais relevantes.

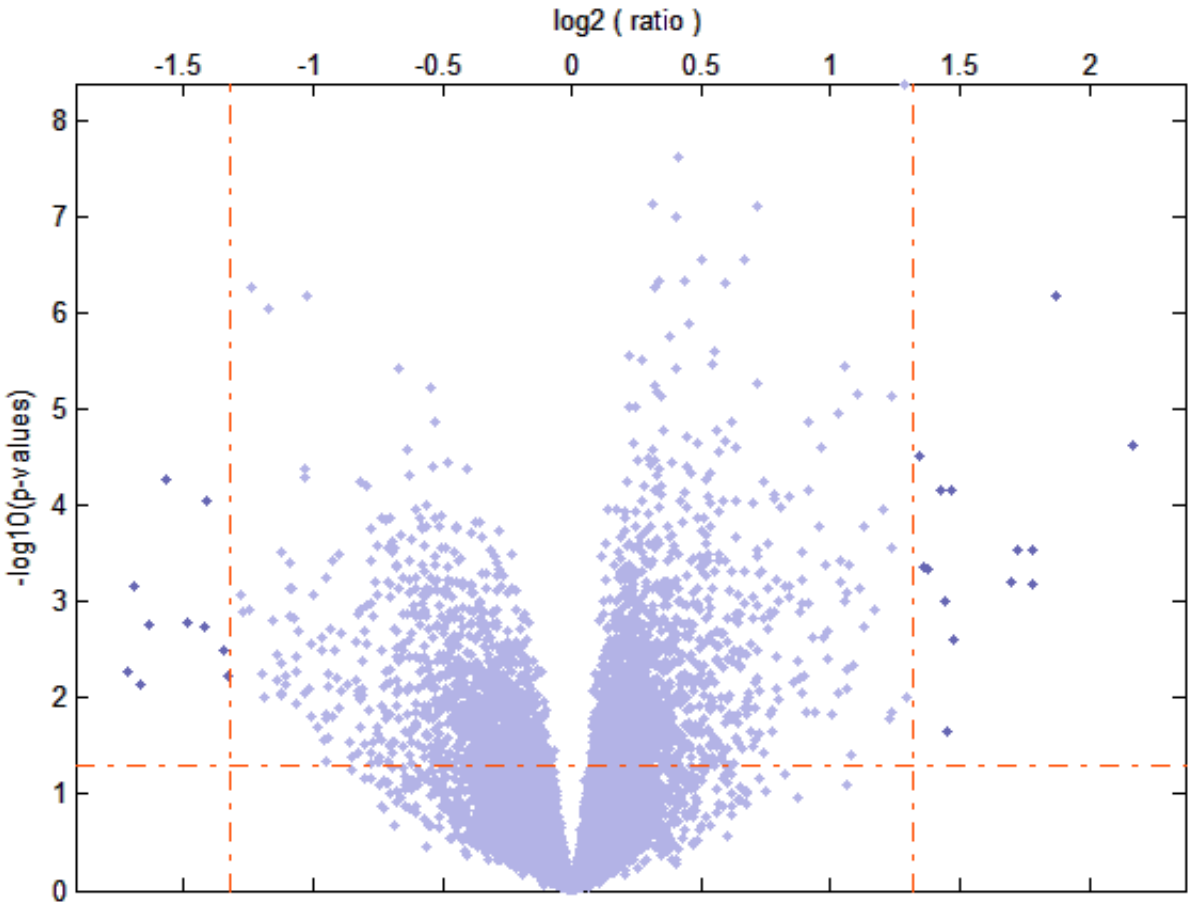
Esse método foi utilizado para chegar ao resultado final do *Volcano Plot*, sendo um passo inicial para ele.

2.2.2 *Volcano Plot*

Esse método utiliza como base o *t-test*, também com dados de p-valor, porém existe uma diferença entre ambos. O Volcano Plot adiciona o *fold change* no seu ajuste, sendo esse termo a medida da quantidade de mudanças que devem ser feitas do valor inicial até o final. Muitas ferramentas matemáticas são normalmente utilizadas para esse tipo de técnica, como por exemplo o MatLab, pois os gráficos gerados são muito visuais e eles permitem

ajustes dos parâmetros através da interface, facilitando muito a seleção das expressões gênicas que se deseja utilizar. Um exemplo pode ser visto na Figura 4, onde foi feito uma exemplo com a base citada anteriormente, de 133 pacientes.

Figura 4 – Exemplo de *Volcano Plot*



Os eixos do *Volcano Plot* são gerados na escala de log, nos quais cada ponto representa a expressão de um gene com base nas duas condições, o p-valor e o *fold change*. O eixo x representa o *fold change*, já o eixo y demonstra a expressão do *t-test*.

Tanto o *t-test* quanto o *Volcano Plot* são métodos muito utilizados em bioinformática, para a seleção de características. Por isso a escolha desses dois métodos para a implementação no presente projeto. Grande parte dos artigos presentes no estado da arte utilizam esses métodos, o que os torna métodos clássicos para esse tipo de problema.

2.2.3 Modelos de Regressões

Dois modelos de regressões foram utilizados neste trabalho. O primeiro deles é o *stepwise regression* e o segundo é o *stepwise generalized linear model*.

A primeira técnica utiliza o *t-test* para fazer a separação dos dados baseados na saída esperada do classificador, gerando dessa forma o hiperplano que seleciona as

principais características para isso. Todas as variáveis que possuem um p-valor razoável para o hiperplano em questão serão adicionadas ao modelo, sendo ajustadas com o p-valor em questão. As demais irão receber um valor 0, indicando que a característica não é boa para o modelo apresentado.

A segunda técnica faz basicamente a mesma coisa, porém ela utiliza uma técnica multivariada da estatística para o ajuste do modelo, que é o *Generalized Linear Model*, mais conhecido na literatura como GLM. Esse modelo trata cada característica de forma separada, e verifica sua correlação com cada uma das demais, individualmente e quando aplicada ao modelo, estabelecendo, a partir de um p-valor de corte, quais características serão inseridas ou não no modelo. Um grande problema encontrado com esse método é o tempo de processamento. Dependendo do tamanho da base de dados fica inviável aplicar a todo o conjunto.

Neste trabalho, foram utilizados os resultados dos modelos de seleção de características para diminuir a dimensionalidade dos dados utilizados. Dois conjuntos foram gerados, um utilizando o modelo *stepwise regression* e posteriormente outro com a aplicação do *stepwise GLM*.

Esses métodos são mais utilizados em problemas de inteligência computacional e acreditamos que poderiam ser aplicados em bioinformática para melhorar os resultados obtidos. Como existem cerca de 22283 expressões gênicas diferentes nas bases de dados para o câncer de mama, é preciso ter cuidado com a quantidade de expressões que serão selecionadas em ambos os métodos, pois podem ser adicionadas ao modelo muitas características, algo que é financeiramente ruim, devido ao custo para a análise dos genes de um indivíduo.

2.3 Técnicas de Classificação

Nessa seção são apresentadas algumas informações teóricas com relação às principais técnicas de classificação implementadas no presente trabalho. Foram utilizados diversos métodos e os principais foram estes.

2.3.1 Redes Neurais Artificiais *Multi-Layer Perceptron* (RNA MLP)

As redes neurais artificiais são um algoritmo construído com base no comportamento biológico dos neurônios presentes no cérebro humano. De maneira simplificada a rede neural conecta uma entrada e uma saída podendo ter uma ou mais camadas intermediárias que deverão ser ajustadas nas etapas de treinamento. Esse tipo de rede é conhecido como *Multi-Layer Perceptron* (MLP). Os dados utilizados na entrada, para os problemas ligados ao câncer de mama e que utilizam dados de *microarray* deverão ser as expressões gênicas. Os dados de saída serão valores binários, onde um dos valores indica RD e o outro indica

PCR. O objetivo na etapa de treinamento é ajustar a rede para que ela consiga identificar novos dados apresentados, gerando uma saída com a classificação dos novos indivíduos. Uma rede pode estar completamente conectada de forma que cada neurônio esteja ligado a todos os neurônios da próxima camada, ou parcialmente conectada de maneira que cada neurônio esteja ligado a um número especifico de neurônios da próxima camada (GARDNER; DORLING, 1998).

Os parâmetros que devem ser ajustados nas RNAs MLP são:

- Quantidade de neurônios
- Algoritmo de aprendizagem
- Função de transferência
- Número de Camadas intermediárias
- Topologia da rede

O principal parâmetro para uma RNA MLP é o algoritmo de treinamento. Ele irá ajustar todo o funcionamento da rede, definindo os pesos em cada camada para que a saída seja o mais próximo da realidade. Quanto melhor for o ajuste dos pesos para o problema, melhor será a qualidade do classificador, levando em consideração casos de *overfitting*, melhorando assim sua acurácia, sensibilidade e especificidade. Quanto melhor forem esses valores, maior a chance de acerto na classificação. Isso é muito importante para diversos classificadores, o que leva a busca de bons valores para essas características. Existem diversos algoritmos para treinamento da rede. Os mais utilizados para classificação em bioinformática são:

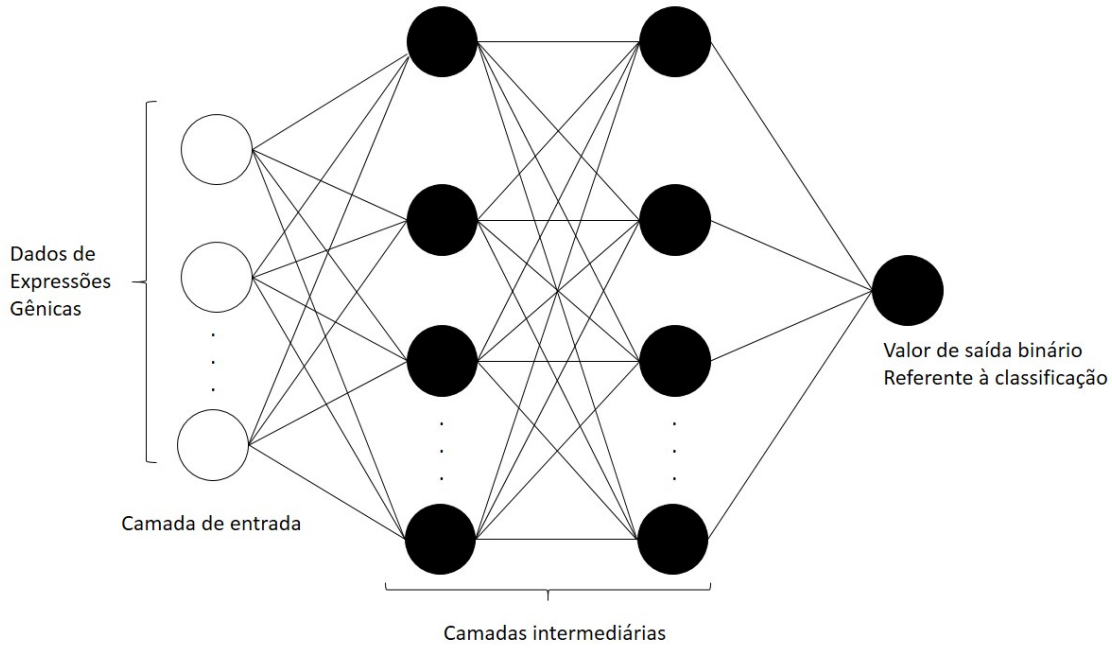
- Backpropagation
- Levenberg-Marquardt
- Conjugate Gradient
- Conjugate Gradient with Powell/Beale Restarts
- Fletcher-Powell Conjugate Gradient
- Polak Ribiere Conjugate Gradient

Um exemplo de estrutura de uma RNA MLP pode ser visto na Figura 5.

2.3.2 Particle Swarm Optimization com clustering

O algoritmo de otimização por enxame de partículas é inspirado no comportamento de bandos de pássaros. A busca por alimentos e a interação entre as aves ao longo do voo são referências para um mecanismo de otimização, sendo que a área sobrevoada é equivalente ao espaço de busca e o fato de encontrar um local com bastante alimento corresponde a encontrar um ótimo local ou até mesmo a solução ótima. Para alcançar o alvo focado, como por exemplo alimentos ou ninhadas, os pássaros fazem uso de suas experiências e da experiências do próprio bando como um todo. O termo que representa

Figura 5 – Exemplo Rede Neural Artificial MLP



essa experiência ou o histórico de vida de cada partícula é denominado como pBest, em contrapartida, o gBest representa o conhecimento do bando todo. Em linhas gerais, essa técnica utiliza uma população de partículas, sendo que cada uma delas representa uma possível solução para o problema em questão.

Cada partícula, ou pássaros, possui um vetor de posição, coordenadas, sendo a sua localização no espaço de busca, e uma velocidade, responsável por guiar as mudanças da posição das partículas durante o "voo", ou seja, durante a execução do processo. Ao longo desse processo, as posições e velocidades serão modificadas dinamicamente com base nas experiências do conjunto como um todo. O algoritmo base pode ser visto na Figura 6. Nele a variável gBest indica o melhor valor global, sendo ele utilizado como referência para o ajuste das partícula. Já o pBest é o valor de cada partícula, que será ajustado baseado na variável indicativa de velocidade e na direção do gBest.

No problema apresentado, para classificação dos indivíduos, o algoritmo foi ajustado para não utilizar os dados dos melhores globais, chamado de gBest, para que cada partícula possa convergir para um melhor local, gerando assim os clusters necessários para fazer a posterior classificação. O PSO implementado utiliza a distância euclidiana para fazer a classificação de cada paciente. Serão gerados dois clusters, um relativo a saída PCR e outro relativo à saída RD, e então é feito o cálculo da distância euclidiana dos paciente para cada um dos clusters. O que possuir a menor distância será utilizado para associar o paciente ao seu grupo de referência. Se a menor distância for em relação ao cluster referente à saída PCR, então o paciente será do tipo PCR, caso contrário ele será do tipo

Figura 6 – Exemplo de Algoritmo do PSO

Algoritmo 3: Algoritmo PSO

```
[x] = PS(max_it, AC_1, AC_2, v_max, v_min, N)
inicializa x (usualmente inicializado randomicamente)
inicializa Δx_i (randomicamente, Δx_i ∈ [v_min, v_max])
t ⇐ 1
while t < max_it do
    for i = 1 até N do
        if f(x_i) > g(p_i) then
            p_i = x_i
        end
        g = i
        for j = indice de vizinhos do
            if g(p_i) > g(p_g) then
                g = j
            end
        end
        Δx_i ⇐ Δx_i +_1 ⊗(p_i − x_i) +_2 ⊗(p_g − x_i)
        Δx_i ∈ [v_min, v_max]
        x_i ⇐ x_i + Δx_i
    end
    t ⇐ t + 1
end
```

RD. Feito isso deverá ser analisado o erro de classificação de cada paciente.

2.3.3 Extreme Learning Machine

Essa metodologia, também chamada de Máquina de Aprendizado Extremo (NTU, 2013), é uma alternativa às redes neurais, que buscam melhorar o tempo de processamento para convergência dos pesos.

O algoritmo de Extreme Learning Machine utiliza um formato de treinamento que não necessariamente irá ajustar todos os pesos o tempo todo, podendo convergir bem mais rápido. Além disso ela possui diversos algoritmos de otimização que podem ser utilizados.

Esse método trabalha com duas camadas básicas, uma de entrada e outra oculta, mais simplificada. Os pesos entre as camadas de entrada e oculta são aleatoriamente escolhidos, e entre as camadas oculta e de saída, são determinados analiticamente. Devido à sua velocidade de aprendizagem diversos autores têm aplicado a rede ELM a uma grande quantidade de problemas de classificação e identificação de padrões.

Devido a isso esse algoritmo foi utilizado no presente trabalho, baseado em bibliotecas para o MatLab, sendo mais uma possibilidade de classificador baseado em dados de expressões gênicas.

3 Metodologia

O presente capítulo trata a metodologia utilizada no projeto, para inicialmente tratar a base de dados, depois implementar os algoritmos utilizados e posteriormente classificar os pacientes. Cada etapa da metodologia possui uma grande importância, sendo a principal delas a seleção das expressões gênicas a serem utilizadas.

3.1 Base de Dados

A base de dados utilizada foi proposta em Itoh et al. (2014). Quando o trabalho foi iniciado ela possuía 133 pacientes, cada um com 22283 expressões gênicas. Ao longo do trabalho a base foi incrementada, passando a ter mais de 400 pacientes, dos quais 305 fazem parte dos três subtipos do câncer de mama não basais, e foram selecionados para trabalhar inicialmente devido ao seu comportamento mais estável, apresentando uma distribuição quasi binomial, diferentemente dos basais, que trabalham com uma distribuição dos dados mais complexa, tornando a tarefa de lidar com os mesmos mais difícil. Após os testes com os não basais, foram executados os mesmos procedimentos que serão descritos ao longo do trabalho, porém com os basais. É necessário aplicar diversos testes estatísticos para encontrar uma distribuição razoável para o subtipo basal.

A base mais atual está separada no agrupamento não basal, contendo 186 pacientes para treinamento e 119 para validação. No treinamento existem 173 pacientes do tipo RD e 13 PCR. Já na validação temos 99 RD e 20 PCR. Além deles, temos também os pacientes do agrupamento basal. No treinamento existem 75 RD e 35 PCR. Já na validação temos 19 RD e 9 PCR. Os dados estão representados na Tabela 1.

Tabela 1 – Separação dos dados utilizados.

	Número de Indivíduos RD	Número de Indivíduos PCR
Dados Não Basais de Treinamento	173	13
Dados Não Basais de Validação	99	20
Dados Basais de Treinamento	75	35
Dados Basais de Validação	19	9

Isso justifica a necessidade de se implementar bons classificadores para o problema, pois a maior parte dos pacientes são RD, e poderiam não passar pelo desgaste da quimioterapia neoadjuvante caso um classificador com elevada taxa de acerto e confiável seja obtido. Alguns tratamentos precisam ser feitos para trabalhar com os dados, pois existe um desbalanceamento das classes, e podem impactar na resposta dos classificadores. Assim

novos ajustes poderão ser feitos para trabalhos futuros, visando melhorar os resultados obtidos.

3.2 Seleção das Expressões Gênicas

Os métodos utilizados para a seleção foram os já citados no capítulo anterior. Nas próximas subseções serão informados como os dados foram selecionados.

3.2.1 *Volcano Plot*

O *Volcano Plot* utiliza o teste t como sua base, que por sua vez foi utilizado com p-valor menor que 0.05, baseado em diversos artigos da literatura, apresentados em capítulos anteriores, que buscam um nível de confiança de 95% e para que fosse aplicado no *Volcano Plot*. (CUI; CHURCHILL, 2003)

A saída do *T Test* foi utilizada no *Volcano Plot*, para encontrar expressões gênicas que possuem um p-valor limite de 0.05 e um *fold change* padrão, de valor 2.(CUI; CHURCHILL, 2003).

Quando esse método foi utilizado a base de dados ainda possuía apenas 133 pacientes, e foram selecionados 31 genes ao se aplicar o *Volcano Plot*, conforme apresentado por Hess et al. (2006). Esses dados foram testados para os algoritmos de redes neurais artificiais e o PSO com clusterização.

Conforme informado anteriormente foram selecionados 31 genes que correspondem a essa configuração. Eles serviram de entrada para os algoritmos de classificação.

3.2.2 Modelos de Regressão

Ambos os modelos de regressão possuem a mesma parametrização, porém um é o modelo linear puro e o outro possui o algoritmo GLM associado, sendo o primeiro univariado e o segundo multivariado.

No primeiro caso foram obtidos 110 expressões gênicas. Já no modelo GLM foram obtidas 151.

Durante a execução dos métodos de seleção observamos que as expressões gênicas obtidas por cada um são bem diferentes, e quase não ocorre interseção entre elas.

3.3 Classificadores

A partir da seleção de genes nas bases de dados, foi reduzido o tamanho das entradas, utilizando um número menor de dimensões. Foram criados diversos classificadores

utilizando bibliotecas do MatLab. Seis baseados em RNAs MLP, um baseado em PSO com clustering e outro baseado no algoritmo de *Extreme Learning Machine*.

3.3.1 Configuração das Redes Neurais Artificiais

As seis redes neurais artificiais criadas possuem 20 nodos, valor com melhores resultados obtido empiricamente, na camada intermediária, e os algoritmos de treinamento foram escolhidos baseados nos padrões existentes na literatura para classificação. Eles foram os seguintes:

- Backpropagation
- Levenberg-Marquardt
- Conjugate Gradient
- Conjugate Gradient with Powell/Beale Restarts
- Fletcher-Powell Conjugate Gradient
- Polak Ribiere Conjugate Gradient

Para cada algoritmo de treinamento foram utilizados os genes selecionados pelos métodos supracitados, sendo eles *Volcano Plot*, *Stepwise Regression* e *Stepwise GLM*, que obtiveram conjuntos de 31, 110 e 151 expressões gênicas respectivamente. Posteriormente foi feito um teste com cada um dos conjuntos encontrados.

3.3.2 Configurações do PSO com Clustering

O algoritmo de PSO foi ajustado para possuir 200 partículas e 2 clusters, cada um representando um possível valor para a saída. A função objetivo utilizada foi a distância Euclidiana, onde a cada iteração os possíveis clusters são verificados com relação aos dados, para tentar encontrar o que separa melhor os dois grupos de classificação.

Os demais parâmetros do algoritmo foram mantidos conforme os padrões da literatura, que inclusive apresentaram os melhores valores nos resultados.

3.3.3 Configurações do Extreme Learning Machine

O algoritmo de *Extreme Learning Machine* foi ajustado para o tipo classificação, a configuração para dois tipos de resposta, que seria o PCR e RD, e conforme a saída esperada ele faz um ajuste automático para tentar classificar em dois tipos distintos apenas, baseado na sua configuração. A parametrização foi a seguinte:

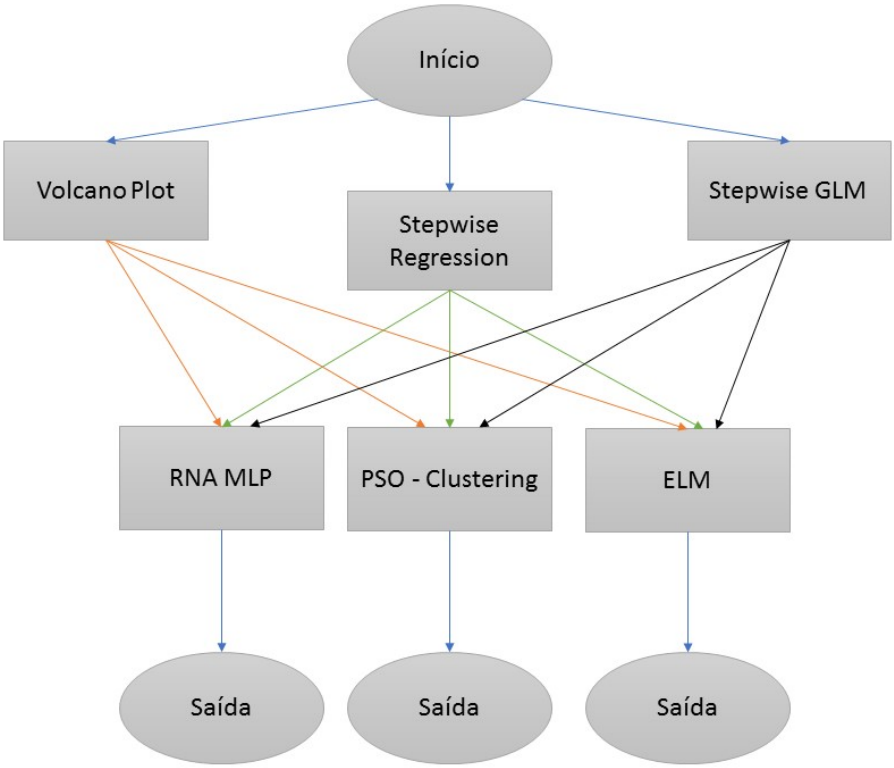
- 75 neurônios na camada escondida.
- 2 agrupamentos: PCR e RD (respostas 1 e 0).
- As funções de ativação utilizadas foram Sigmoidal e Hardlim.

A seleção de 75 neurônios foi efetuada baseado em resultados de teste empíricos. Já o número de agrupamentos é feito considerando a quantidade de classes a serem classificadas no problema. As funções de ativação foram escolhidas baseado nos melhores valores obtidos de forma empírica para a classificação.

Os dados contendo as 151 expressões gênicas obtidas a partir do modelo de regressão baseado na técnica de GLM apresentaram melhores resultados, conforme será apresentado posteriormente.

A figura Figura 7 apresenta um diagrama básico do funcionamento do sistema criado, onde a primeira etapa compreende a seleção das expressões gênicas e a segunda etapa a utilização dos algoritmos de classificação, inclusive o *ELM*.

Figura 7 – Diagrama básico do funcionamento dos algoritmos.



4 Resultados e Discussões

Os resultados obtidos são sumarizados na Tabela 2. Eles serão discutidos separadamente, porém é interessante ter uma visão do todo para que os detalhes sejam entendidos melhor. Podemos perceber que o *ELM (Extreme Learning Machine)* possui o melhor resultado para todos os conjuntos de expressões gênicas.

Inicialmente foi feito um tratamento nos dados para selecionar os conjuntos de expressões gênicas a serem trabalhados. Foram utilizadas três formas diferentes de seleção, duas univariadas (*Volcano Plot* e *Stepwise Regression*) e uma multivariada (*Stepwise GLM*), cujos resultados serão discutidos posteriormente.

4.1 Aplicação de Técnicas de Seleção

4.1.1 *Volcano Plot*

O primeiro método utilizado foi o *Volcano Plot*. Essa técnica utiliza um valor de corte baseado no *p-values* que foi de 95%, e um *fold change* de 2.

O resultado obtido com essa técnica foi o da Figura 4, apresentada na descrição do método. Foram obtidas 31 expressões gênicas para essa técnica, indicando a primeira linha de resultados da Tabela 2.

4.1.2 *Modelo de Regressão: Stepwise Regression*

O segundo método utilizado foi o *Stepwise Regression*. O modelo trabalha conforme descrito nas seções anteriores, e foram selecionadas 110 expressões gênicas relevantes com esse método. Os resultados para os classificadores executados com essa base estão apresentados na segunda linha da Tabela 2.

Tabela 2 – Porcentagem de acerto para cada um dos algoritmos, de acordo com o conjunto de expressões gênicas utilizado

	RNAs	PSO com Clustering	ELM
Volcano Plot - 31	52%	48%	68%
Stepwise Regression - 110	58%	52%	71%
Stepwise GLM 151	59%	62%	83%

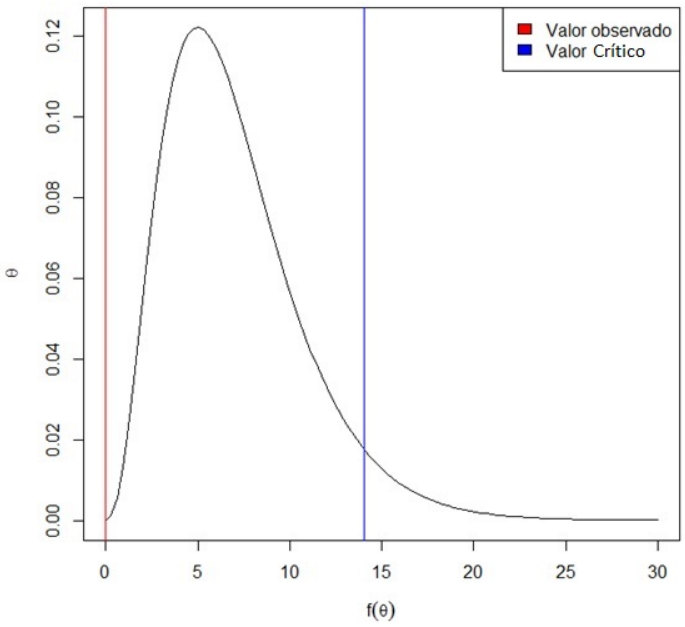
4.1.3 *Modelo de Regressão: Stepwise GLM*

O terceiro método para seleção de características foi o *Stepwise GLM*, que utiliza o mesmo formato do anterior, porém aplica a técnica estatística GLM para a verificação das características mais relevantes, que é multivariada. Assim uma validação de cada expressão gênicas com relação a todas as demais é feita.

Para que esse método seja utilizado de forma correta é necessário fazer uma verificação da distribuição estatística dos dados. Para isso algumas delas foram realizadas e o resultado encontrado foi o da distribuição quasibinomial, conforme evidenciado pela Figura 8. Podemos observar que o valor obtido é menor que o valor crítico para a distribuição, indicando que o ajuste está correto e assim a distribuição quasibinomial é compatível com a distribuição encontrada. Essa figura apresenta uma análise do ajuste dessa distribuição, e podemos ver que o valor observado é menor que o ponto crítico, indicando que essa distribuição é bem ajustada.

Os resultados para os classificadores utilizando essa técnica estão apresentados na terceira linha da Tabela 2.

Figura 8 – Gráfico para análise do ajuste. A reta vermelha apresenta o valor observado e a azul o valor de corte para identificar se o ajuste está correto ou não. Como pode ser visto no gráfico o ajuste está adequado.



4.2 Classificadores

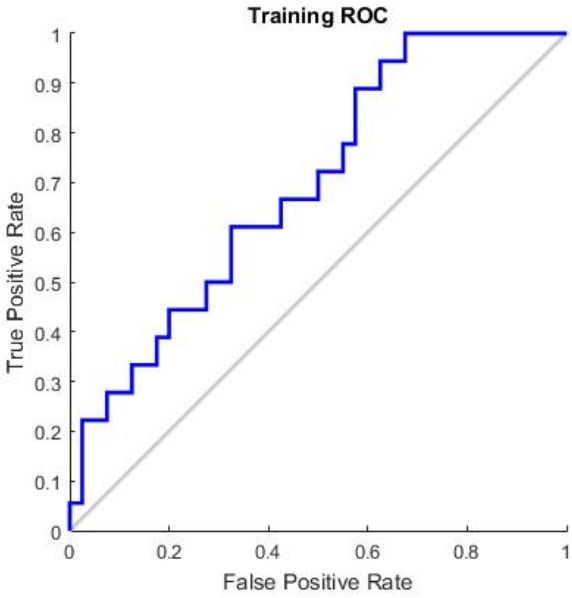
4.2.1 Experimentos feitos com Redes Neurais Artificiais

As redes neurais são vastamente utilizadas em problemas de classificação e por isso foram aplicadas no presente trabalho, na tentativa de obter um bom resultado com uma acurácia melhor que a existente na literatura.

Inicialmente o classificador com redes neurais foi utilizado para a base de 31 expressões gênicas, e foram geradas curvas ROC para avaliar o seu desempenho.

Conforme podemos observar na Figura 9, o comportamento não foi muito bom para o treinamento, obtendo um valor na área abaixo da curva variando de 0.42 a 0.50. Na Figura 10 observamos que a validação apresentou um resultado melhor para um teste inicial, chegando a cerca de 0.72 a área abaixo da curva, para os melhores testes. Já na curva ROC geral, presente na Figura 10, observamos que a qualidade do classificador foi de aproximadamente 50%, sendo na verdade 0.52 o valor da área abaixo da curva ROC. Esse tipo de curva possui uma área total de 1, e seus eixos representam a sensibilidade e a especificidade do classificador. Sendo assim, o classificador obtido não apresentou um bom resultado aparente, ficando com uma média de 52% de acerto para os 50 testes executados.

Figura 9 – Curva ROC de treinamento



O valor obtido variando o tipo de treinamento foi aproximadamente o mesmo, não existindo nem mesmo diferença estatística entre os classificadores, baseado na tabela ANOVA com confiança de 95%.

Figura 10 – Curva ROC de validação

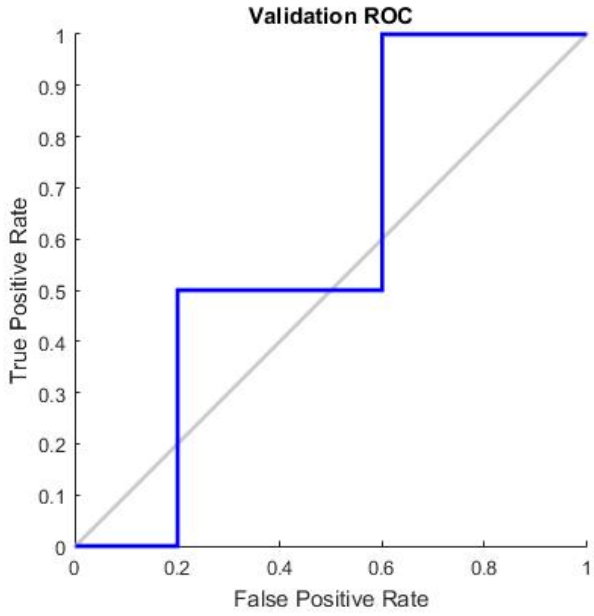
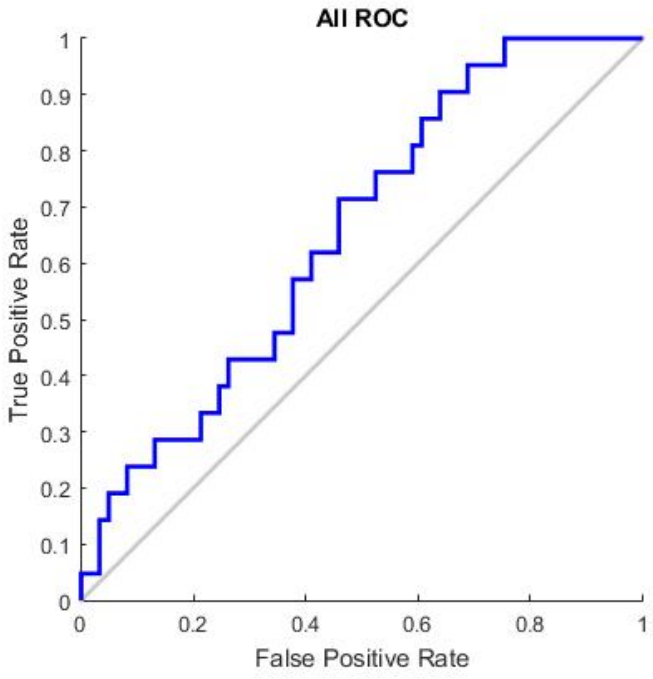


Figura 11 – Curva ROC geral



Fazendo o mesmo processo para as bases de dados de 110 e 151 pacientes foram gerados resultados um pouco melhores, chegando a 58% de acerto, porém ainda é um resultado aparentemente ruim.

Devido a isso, fizemos um novo teste para avaliar se existe diferença estatística entre

testes feitos variando a quantidade de expressões gênicas na entrada das redes neurais, ou seja para avaliar se alterando o número de pacientes o resultado do classificador seria diferente. Foi considerada a base de 110 expressões gênicas para isso, pois era a mais relevante no momento dos testes com redes neurais.

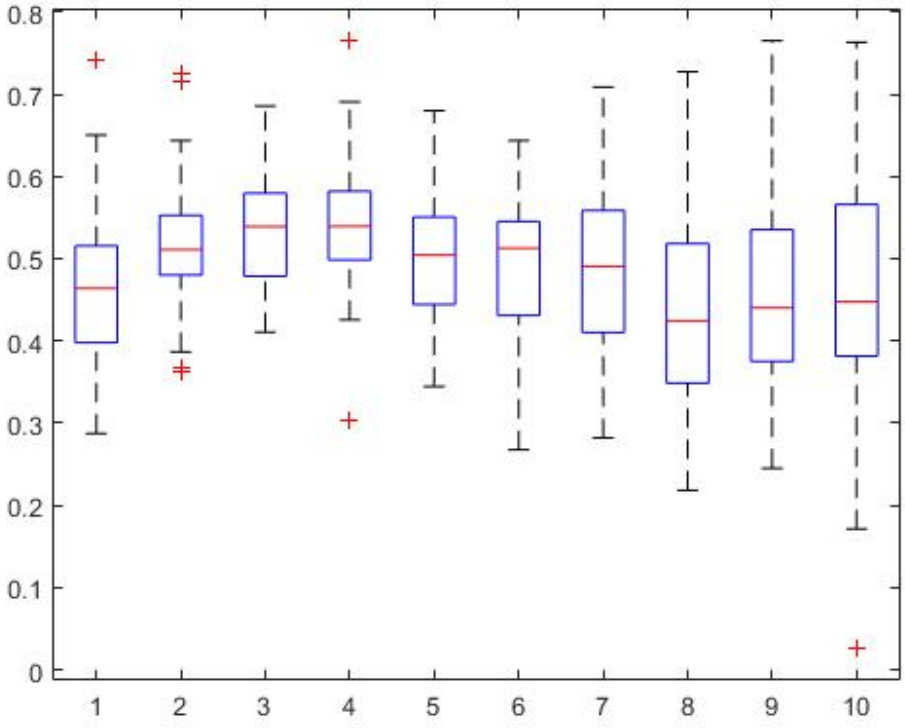
Para isso foi utilizada a base de dados com 110 expressões gênicas, iniciando com 20 delas de forma aleatória e aumentando esse valor de 10 em 10 até chegar nas 110 finais, gerando uma variação na proporção de expressões gênicas utilizadas no treinamento. Os dados de resposta utilizados para fazer essa análise foram os da área acima da curva ROC, que indicam a taxa de erro do classificador, para que uma comparação entre as redes neurais fosse avaliada.

Para fazer essa análise é necessário aplicar a técnica ANOVA (Análise de variância), e assim suas premissas precisam ser atendidas.

Podemos verificar pela tabela ANOVA que apresentou um resultado de 3.32e-07, que existe diferença estatística, porém ao analisar o teste de Tukey na Figura 16 vemos que o mesmo não corrobora completamente com o resultado da ANOVA, pois existem algumas interseções entre alguns conjuntos de teste. No teste de Tukey ocorrem algumas interseções entre os testes de melhor resultado e outros que foram piores, ficando apenas 4 grupos estatisticamente diferentes do resultado com 110 expressões gênicas, que possui um erro variando entre 56% e 62%. Todas as premissas da tabela ANOVA foram atendidas, conforme apresentado na Figura 12, onde temos o boxplot criado variando o número expressões gênicas e verificando o erro médio para cada conformação do classificador. Na Figura 13 temos o gráfico dos resíduos, na Figura 14 temos a homocedasticidade e por fim o QQplot na Figura 15.

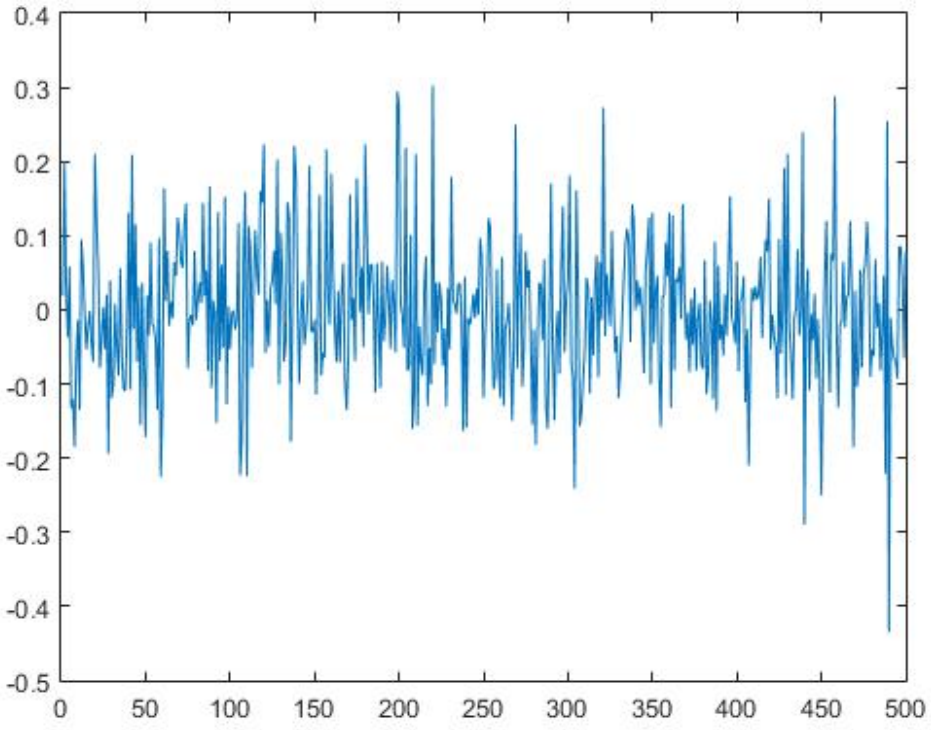
Podemos ver pela Figura 12 que as médias estão bem próximas, um evidência inicial de que provavelmente não existe diferença estatística entre os classificadores.

Figura 12 – Boxplot - Indicando a variação da média dos erros de classificação para cada um dos conjuntos de expressões gênicas.



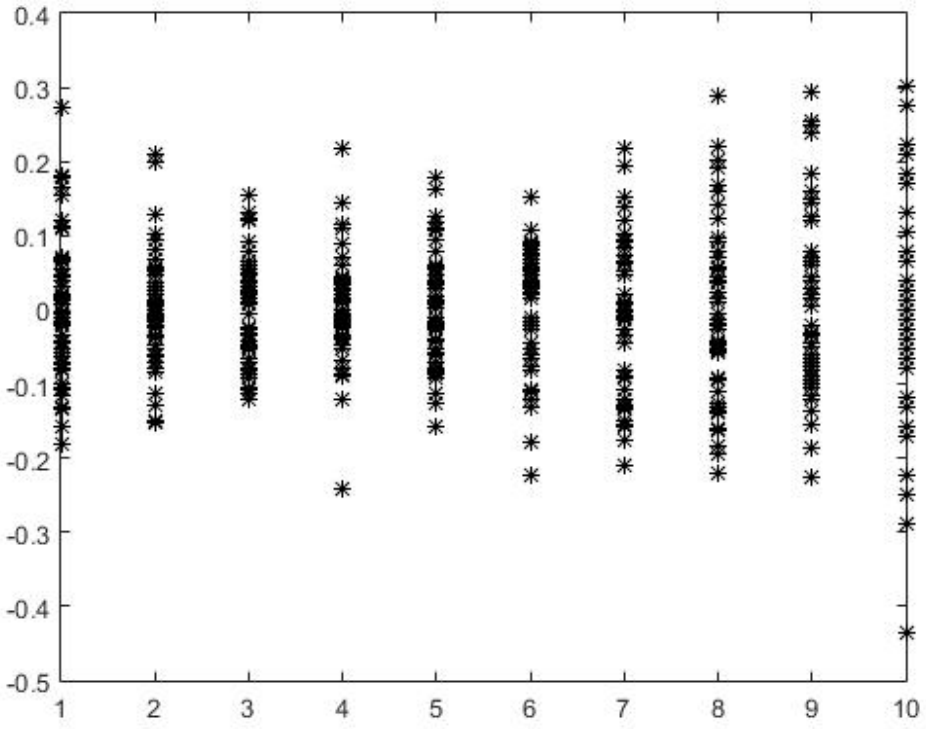
Uma das premissas da ANOVA é a verificação dos resíduos, que precisam apresentar uma distribuição sem padrões, para indicar aleatoriedade dos dados. Podemos ver que isso ocorre, baseado na Figura 13.

Figura 13 – Plot dos resíduos - técnica utilizada para verificar se existe algum padrão entre os resultados dos resíduos.



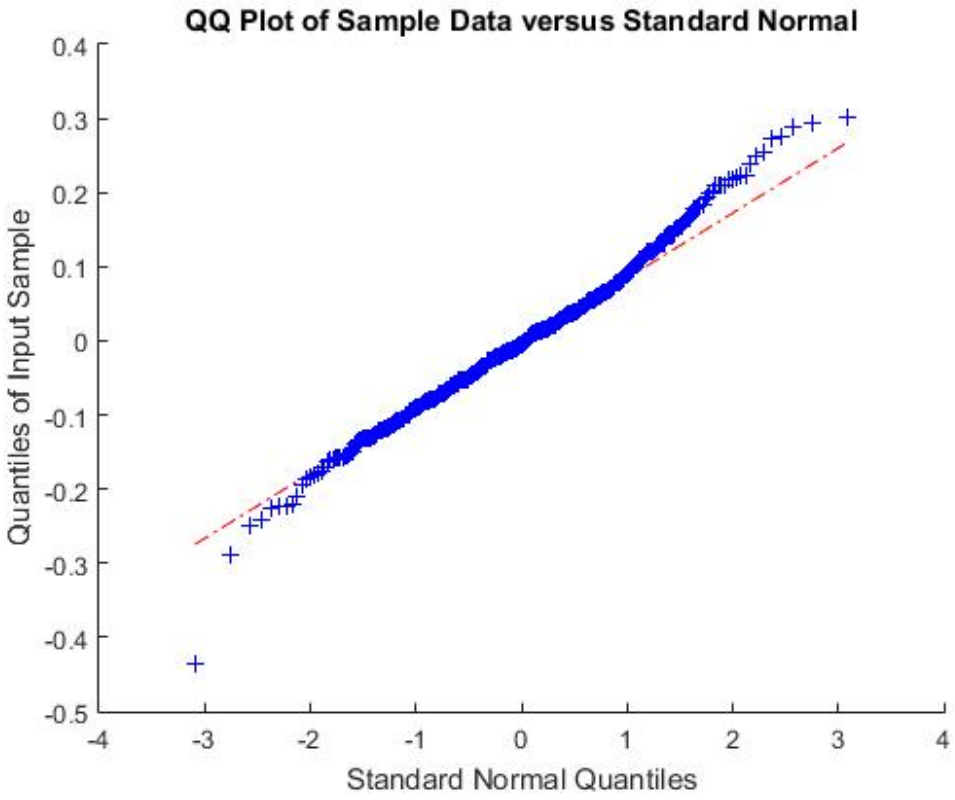
O mesmo comportamento deve acontecer para a homocedasticidade, onde é verificada a variância constante dos erros experimentais. O ideal é que não ocorram padrões nos pontos referentes a esse item, conforme podemos evidenciar pela Figura 14. Os resíduos devem apresentar a mesma variância para cada observação e assim será avaliado o conteúdo informacional dos resíduos. A partir disso é analisada a dispersão dos resíduos em torno da sua média, obtendo o gráfico da Figura 14.

Figura 14 – Homocedasticidade - Técnica utilizada para verificar padrões nos ajustes dos resíduos. Uma das premissas da ANOVA



E por último, é necessário verificar se o QQplot apresenta grandes variações na região mais central da distribuição, podendo variar mais apenas no começo e no final. Isso pode ser evidenciado na Figura 15

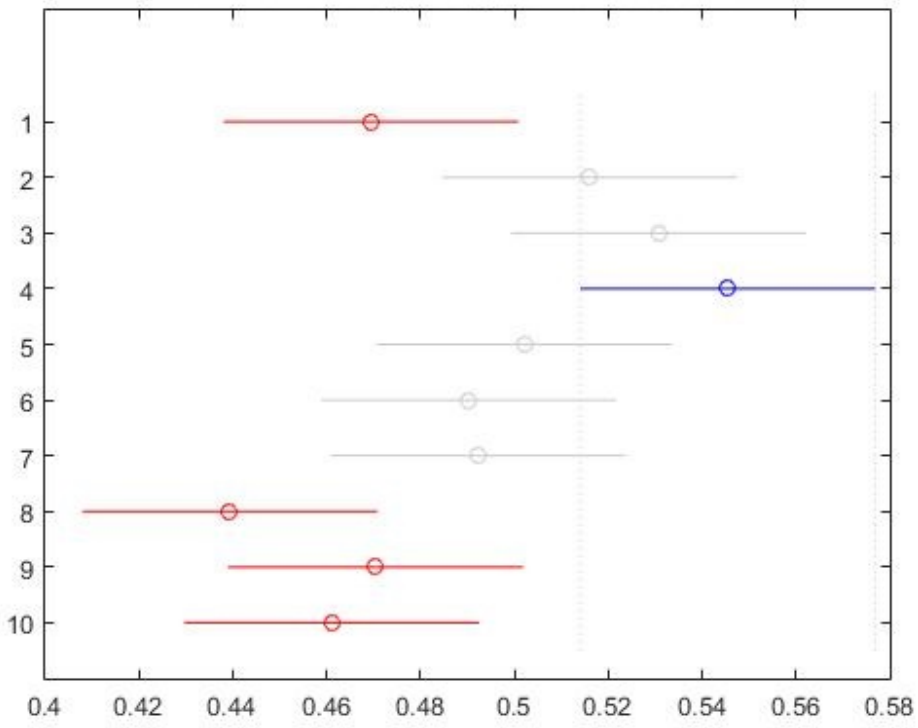
Figura 15 – QQplot - Indica se existe variação com relação a regressão aplicada e os dados, podendo variar apenas nas pontas.



Como todas as premissas foram verificadas, a tabela ANOVA pode ser gerada para validar se existe ou não diferença estatística entre os classificadores e assim corroborar com o resultado apresentado anteriormente. O resultado obtido para esse teste foi de 3.32e-07, indicando que aparenta existir diferença estatística, que deverá ser confirmada pelo teste de Tukey.

E para confirmar se existe ou não a diferença estatística, com mais precisão, foi executado um teste de Tukey que verifica a sobreposição dos resultados para todas as distribuições. Podemos ver que elas não são estatisticamente diferentes, conforme mostrado na Figura 16.

Figura 16 – Teste de Tukey - Verifica diferença estatística utilizando os dados da ANOVA, para indicar sobreposição dos resultados.



A partir dos testes feitos, pode ser verificado que a quantidade de expressões gênicas não afeta no resultado para redes neurais, e não diferencia tanto os grupos, assim, apenas 4 foram estatisticamente diferentes dos demais, mas ainda ocorre sobreposição de seus vizinhos com os outros. Devido a isso novas bases de dados devem ser testadas e até mesmo agrupadas a atual para uma nova verificação baseada nessa quantidade de pacientes, pois apenas com mais pacientes o algoritmo de redes neurais pode ser melhorado. Testes semelhantes foram feitos para verificar a influência da quantidade de pacientes utilizada no treinamento, e foi verificado que esse número influencia na melhoria da rede, porém não foi possível testar até que ponto essa influência pode gerar melhorias no classificador, pois seria necessário utilizar uma base de dados maior para obter a estabilidade dessa melhoria.

Um detalhe a ser levado em consideração é que apenas os 4 tipos (*Backpropagation*, *Levenberg-Marquardt*, *Conjugate Gradient*, *Conjugate Gradient with Powell/Beale Restarts*) de treinamento com melhor resultado foram considerados para análise estatística dos resultados. Os demais forneceram uma taxa de acerto muito ruim inicialmente e não foram utilizados no trabalho (*Fletcher-Powell Conjugate Gradient*, *Polak Ribiere Conjugate Gradient*).

4.2.2 Experimentos utilizando PSO com Clusterização

O PSO com clusterização apresentou inicialmente uma varição no resultado quando a quantidade de expressões gênicas utilizadas era aumentado. Inicialmente foi feito um teste com 31 expressões gênicas, conforme a seleção pelo *Volcano Plot*. O resultado foi de um erro aproximado de 52%. Com o aumento para 110 genes o resultado já passou para 42% de erro. Já para a utilização de 151 expressões o erro reduziu para 38%. Esses valores foram obtidos aplicando a média dos resultados para testes cruzados.

Esses resultados iniciais geraram a suspeita de que a quantidade de expressões gênicas utilizadas para esse método poderiam afetar a resposta. Com isso uma nova análise estatística multivariada foi executada para avaliar se o modelo é bem ajustado e se sofre influência da quantidade de expressões gênicas.

Os dados de erros obtidos para cada um dos algoritmos implementados foram utilizados para verificar se o modelo é bem ajustado ou não. Esses dados correspondem a 30 execuções de cada uma das 4 redes neurais e do PSO com clusterização. Todos eles foram testados com 31, 110 e 151 expressões gênicas, conforme informado anteriormente. Feito isso, foi aplicado o algoritmo *Generalized Linear Model*(GLM) com distribuição quasibinomial, que foi a aplicável ao problema conforme a Figura 8, que representa a saída da execução do ajuste do modelo utilizando a ferramenta R. Na Figura 17, vemos que os p-valores obtidos para cada ajuste do modelo, na parte do dos coeficientes, apresentam valores razoáveis para a aceitação do modelo, utilizando a distribuição quasibinomial. Utilizando o tipo de algoritmo como fato (foram utilizados os dados dos erros para as redes neurais e para o PSO, para verificar a influência das quantidades de expressões gênicas no resultado), e tentando explicar o erro obtido baseado na quantidade de expressões gênicas e no algoritmo utilizado.

O resultado obtido está representado na Figura 17. Com base nos *p-values* verificamos que o modelo é bem ajustado, ou seja, explica bem os valores das médias. Vemos também na figura página 24, pela análise do ajuste do modelo, que ele é realmente bem ajustado, pois o valor observado, em vermelho, aparece antes do valor crítico, não rejeitando a hipótese nula de que ele não é ajustado.

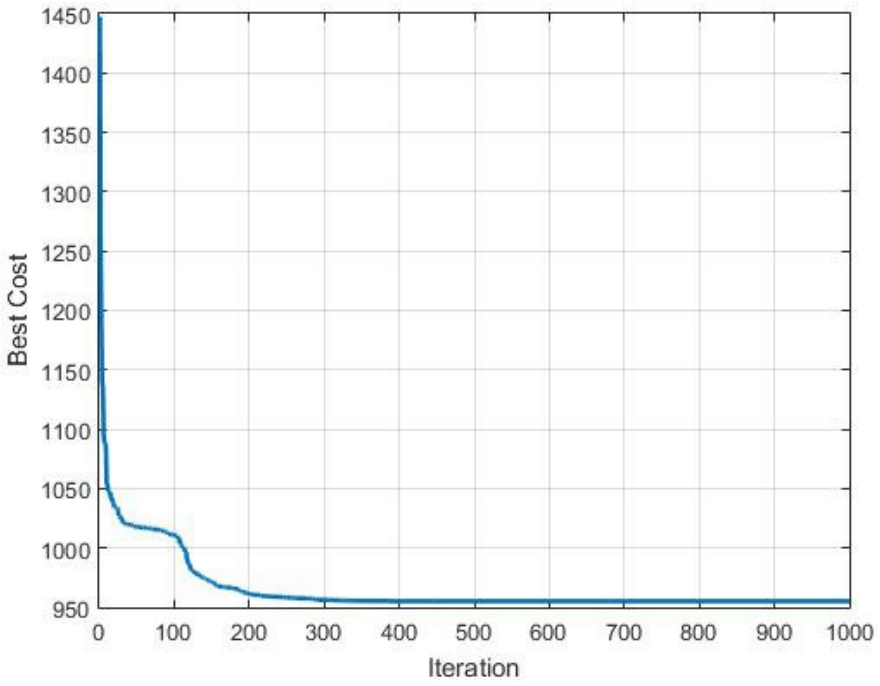
Além disso, pode ser visto pelo coeficiente das expressões, caso o seu número aumente, que o erro de classificação terá uma tendência a diminuir para o PSO, pois são inversamente proporcionais. Porém, utilizar uma grande quantidade de expressões pode inviabilizar o processamento dos algoritmos, devido à quantidade de dados. Assim novas formas de seleção podem ser aplicadas para tentar aumentar o número de expressões, porém sem tornar o processamento inviável. Isso indica que novos algoritmos podem ser utilizados baseados no mesmo princípio, para tentar classificar os pacientes, e seus dados poderão ser utilizados para comparação com os demais.

Figura 17 – GLM quasibinomial

```
glm(formula = MediaErro ~ Expressoes + Algoritmo, family = quasibinomial(),
    data = dados)

Deviance Residuals:
       Min          1Q      Median          3Q         Max
-1.557e-05   2.110e-08   2.110e-08   2.110e-08   1.839e-05

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 321.5676     5.6956   56.46 1.43e-10 ***
Expressoes   -2.3546     0.0697  -33.78 5.16e-09 ***
Algoritmo   -45.3339     0.8039  -56.39 1.45e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasibinomial family taken to be 1.240413e-10)

    Null deviance: 6.5017e+00  on 9  degrees of freedom
Residual deviance: 6.3889e-10  on 7  degrees of freedom
```

Um exemplo do comportamento da função objetivo do PSO pode ser visto na Figura 18, indicando uma boa convergência.

Figura 18 – Comportamento da função objetivo do PSO



Assim, podemos verificar que com uma maior quantidade de expressões gênicas os resultados aparentemente deverão melhorar, porém elas devem ser muito bem selecionadas, para não chegarem a um limite computacional. Executar com todas as expressões, por exemplo, é inviável para o algoritmo de PSO com clusterização, pois gera um tempo de processamento enorme e excede o limite de iterações para execução do processo.

4.2.3 Experimentos feitos com *Extreme Learning Machine (ELM)*

Os experimentos feitos com o algoritmo *ELM* foram os mais satisfatórios até o momento. O melhor resultado encontrado foi utilizando 75 neurônios na camada escondida, 2 grupos para a classificação (PCR e RD) e 151 expressões gênicas para cada paciente como dados de entrada. O resultado obtido para 30 execuções foi uma média de aproximadamente 83% de acerto na validação, baseado em cerca de 95% de acerto no treinamento, sendo um resultado muito promissor. Mesmo não apresentando resultados tão bons com os outros conjuntos de expressões gênicas, o seu resultado foi melhor que o dos demais algoritmos, apresentando cerca de 68% de acerto para 31 expressões gênicas e 71% para o conjunto de 110 expressões. Na Tabela 3 podemos ver a tabela de confusão para o resultado utilizando as 151 expressões gênicas para o algoritmo GLM. Essa tabela indica que existe uma grande taxa de acerto para os pacientes RD, porém a taxa é de apenas 52% para os PCR, indicando uma baixa taxa para esse grupo. É necessário ajustar um pouco os dados de entrada e o algoritmo para que ele melhore essa taxa para os pacientes PCR.

Tabela 3 – Tabela de confusão para os dados não basais.

	RD	PCR
Classificado como RD	87%	13%
Classificado como PCR	48%	52%

Após a obtenção dos melhores resultados com o ELM, o mesmo teste foi aplicado também aos dados basais. Foram selecionadas 138 expressões gênicas para cada paciente, utilizando o algoritmo de *Stepwise GLM*, que apresentou os melhores resultados para seleção. O classificador apresentou resultado médio de 72% de acerto, conforme esperado, devido ao problema de distribuição estatística para os dados não basais. Esses dados ainda precisam ser testados com outras configurações, para encontrar uma distribuição mais adequada para os mesmos. A tabela Tabela 4 apresenta a tabela de confusão para esse teste. Podemos ver que seu resultado com relação à tabela foi razoável, porém o classificador geral apresenta uma baixa taxa de acerto médio ainda.

Tabela 4 – Tabela de confusão para os dados basais.

	RD	PCR
Classificado como RD	76%	24%
Classificado como PCR	29%	71%

Um detalhe interessante observado é o de que os dados basais apresentaram 138 expressões gênicas para o algoritmo de seleção utilizando o método GLM, e os não basais apresentaram 151, porém ao verificar quais eram semelhantes, apenas 1 expressão gênicas existe nos dois conjuntos, indicando que o comportamento dos basais é realmente diferente dos não basais e seu classificador utiliza expressões diferentes para defini-los de forma

mais correta. Isso corrobora com a premissa inicial de que os conjuntos são bem diferentes em termos das distribuições dos dados.

5 Conclusão

No presente trabalho foram aplicadas diversas técnicas de Inteligência Computacional, Computação Evolucionária e Estatística, tanto para a seleção das expressões gênicas, quanto para a implementação de classificadores para pacientes com câncer de mama.

Ainda é preciso melhorar e analisar alguns resultados encontrados, porém o ELM com mais de 80% de acerto já fornece um valor razoável para a continuidade do trabalho. Esse algoritmo possui uma execução extremamente rápida e apresentou o melhor resultado aparente até o momento, sendo um forte candidato para a criação de um preditor mais elaborado.

Novas bases de dados deverão ser utilizadas para tentar melhorar os resultados, e estas ainda precisam ser pesquisadas para verificar qual utilizar. Seria ideal que todas possuam o mesmo formato, para que se possa mesclar e gerar uma única grande base com um conjunto de pacientes bem maior.

As redes neurais não apresentaram um resultado tão bom e devido a isso não seriam bons algoritmos para se continuar tentando melhorar. Acreditamos que o ELM seja a melhor opção para se obter resultados melhores, porém outros algoritmos não utilizados ainda deverão ser avaliados e implementados para comparação com já utilizados.

A principal conclusão até o momento é que o ELM aparenta ser o melhor algoritmo e forneceu os melhores resultados iniciais, mas ainda precisa ser analisado com mais embasamento estatístico, para garantir sua boa qualidade.

Outra conclusão interessante é que os métodos de seleção das expressões gênicas influenciam no resultado dos classificadores. Podemos perceber que o método multivariado *Stepwise GLM* apresentou os melhores resultados para todos os algoritmos, indicando que a seleção dos dados a serem utilizados nos algoritmos para classificação são importantes e podem afetar muito o resultado, chegando a cerca de 13% de melhoria para um mesmo algoritmo.

Os resultados obtidos seguiram a mesma metodologia da literatura, onde se utiliza algoritmos para reduzir a dimensionalidade dos dados e então se aplica classificadores para identificar a resposta de cada indivíduo. A principal diferença é a escolha dos algoritmos de seleção e de classificação.

Podemos verificar que o resultado obtido para o *ELM* foi muito superior aos mencionados anteriormente, na literatura (76% de acerto), para a classe RD, porém ele está desbalanceado para a classe PCR. Alguns ajustes são necessários para que não ocorra esse desbalanceamento. Porém, a classe de maior expressão é a RD, onde cerca de 90%

dos indivíduos se encaixam e o classificador consegue acertar quase 90% dos casos para ela.

Já os demais classificadores não apresentaram o desempenho esperado, sendo piores que os existentes na literatura.

6 Trabalhos Futuros

Para trabalhos futuros, buscamos utilizar mais bases de dados para complementar a atual e tentar obter resultados ainda melhores.

Ainda é necessário fazer uma comparação do ELM com os demais algoritmos e avaliar a sensibilidade e a especificidade do algoritmo, para verificar a qualidade do resultado. Feito isso o próximo passo será tentar criar novos filtros para as expressões gênicas e fazer testes reduzindo e aumentando sua quantidade. Testes estatísticos ainda precisam ser elaborados para testar a qualidade do resultado.

A busca por novos algoritmos de classificação baseados em aprendizado de máquina, computação evolucionária e inteligência computacional também é um dos novos objetivos. Eles deverão também ser testados e comparados com os demais para avaliar a qualidade de cada um. Além disso queremos testar a combinação de algoritmos diferentes para verificar se o resultado fica melhor.

Referências

ANDRE, F. et al. Dna arrays as predictors of efficacy of adjuvant/neoadjuvant chemotherapy in breast cancer patients: current data and issues on study design. **Biochimica et Biophysica Acta (BBA)-Reviews on Cancer**, Elsevier, v. 1766, n. 2, p. 197–204, 2006. Citado na página 3.

BENSMAIL, H.; HAOUDI, A. Postgenomics: proteomics and bioinformatics in cancer research. **BioMed Research International**, Hindawi Publishing Corporation, v. 2003, n. 4, p. 217–230, 2003. Citado na página 3.

BERGAMASCHI, A. et al. Distinct patterns of dna copy number alteration are associated with different clinicopathological features and gene-expression subtypes of breast cancer. **Genes, Chromosomes and Cancer**, Wiley Online Library, v. 45, n. 11, p. 1033–1040, 2006. Citado na página 4.

BUCHHOLZ, T. A. et al. Global gene expression changes during neoadjuvant chemotherapy of human breast cancer: 9: 21 am (14). **The Cancer Journal**, LWW, v. 8, n. 6, p. 488, 2002. Citado na página 3.

CAVALLARO, S. et al. Genomic analysis: toward a new approach in breast cancer management. **Critical reviews in oncology/hematology**, Elsevier, v. 81, n. 3, p. 207–223, 2012. Citado na página 4.

CHIN, K. et al. Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. **Cancer cell**, Elsevier, v. 10, n. 6, p. 529–541, 2006. Citado na página 4.

CHIN, S. F. et al. High-resolution acgh and expression profiling identifies a novel genomic subtype of er negative breast cancer. **Genome biology**, BioMed Central, v. 8, n. 10, p. 1, 2007. Citado na página 4.

CIRQUEIRA, M. B. et al. Subtipos moleculares do câncer de mama. **Femina**, v. 39, n. 10, 2011. Citado 2 vezes nas páginas 4 e 5.

COLOMBO, J.; RAHAL, P. A tecnologia de microarray no estudo do câncer de cabeça e pescoço. **Revista Brasileira de Biociências**, p. 64–72, 2010. Citado 2 vezes nas páginas 3 e 7.

CUI, X.; CHURCHILL, G. A. Statistical tests for differential expression in cdna microarray experiments. **Genome biology**, BioMed Central, v. 4, n. 4, p. 210, 2003. Citado na página 20.

DONAHUE, H. J.; GENETOS, D. C. Genomic approaches in breast cancer research. **Briefings in functional genomics**, Oxford University Press, v. 12, n. 5, p. 391–396, 2013. Citado na página 4.

ELSTON, C. W.; ELLIS, I. O. Pathological prognostic factors in breast cancer. i. the value of histological grade in breast cancer: experience from a large study with long-term follow-up. **Histopathology**, Wiley Online Library, v. 19, n. 5, p. 403–410, 1991. Citado na página 1.

FISHER, B. et al. Effect of preoperative chemotherapy on the outcome of women with operable breast cancer. **Journal of Clinical Oncology**, American Society of Clinical Oncology, v. 16, n. 8, p. 2672–2685, 1998. Citado na página 2.

GARDNER, M. W.; DORLING, S. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. **Atmospheric environment**, Elsevier, v. 32, n. 14, p. 2627–2636, 1998. Citado na página 16.

GONG, Y. et al. Determination of oestrogen-receptor status and erbb2 status of breast carcinoma: a gene-expression profiling study. **The lancet oncology**, Elsevier, v. 8, n. 3, p. 203–211, 2007. Citado na página 3.

GRALOW, J. R. et al. Preoperative therapy in invasive breast cancer: pathologic assessment and systemic therapy issues in operable disease. **Journal of Clinical Oncology**, American Society of Clinical Oncology, v. 26, n. 5, p. 814–819, 2008. Citado na página 2.

HATZIS, C. et al. A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer. **Jama**, American Medical Association, v. 305, n. 18, p. 1873–1881, 2011. Citado na página 12.

HERSCHKOWITZ, J. I. et al. Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors. **Genome biology**, BioMed Central, v. 8, n. 5, p. 1, 2007. Citado na página 6.

HESS, K. R. et al. Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer. **Journal of clinical oncology**, American Society of Clinical Oncology, v. 24, n. 26, p. 4236–4244, 2006. Citado 4 vezes nas páginas 3, 7, 8 e 20.

HOADLEY, K. A. et al. Egfr associated expression profiles vary with breast tumor subtype. **BMC genomics**, BioMed Central, v. 8, n. 1, p. 1, 2007. Citado na página 6.

HORTA, E. G. **Previsores para a eficiência da quimioterapia neoadjuvante no câncer de mama**. Novembro de 2008. 100 p. Dissertação (Mestrado em Ciência da Computação) — Centro Federal de Educação Tecnológica de Minas Gerais, Belo Horizonte, 2008. Citado 3 vezes nas páginas viii, 7 e 8.

HU, Z. et al. The molecular portraits of breast tumors are conserved across microarray platforms. **BMC genomics**, BioMed Central Ltd, v. 7, n. 1, p. 96, 2006. Citado na página 6.

ITOH, M. et al. Estrogen receptor (er) mrna expression and molecular subtype distribution in er-negative/progesterone receptor-positive breast cancers. **Breast cancer research and treatment**, Springer, v. 143, n. 2, p. 403–409, 2014. Citado 2 vezes nas páginas 12 e 19.

JACOBS, C.; SIMOS, D.; CLEMONS, M. Treatment for locally advanced breast cancer: a global challenge, personalized medicine or both? **Current opinion in supportive and palliative care**, LWW, v. 8, n. 1, p. 30–32, 2014. Citado na página 2.

JEMAL, A. et al. Global cancer statistics. **CA: a cancer journal for clinicians**, Wiley Online Library, v. 61, n. 2, p. 69–90, 2011. Citado na página 4.

KESSLER, J. D. et al. A sumoylation-dependent transcriptional subprogram is required for myc-driven tumorigenesis. **Science**, American Association for the Advancement of Science, v. 335, n. 6066, p. 348–353, 2012. Citado na página 6.

KUERER, H. M. et al. Clinical course of breast cancer patients with complete pathologic primary tumor and axillary lymph node response to doxorubicin-based neoadjuvant chemotherapy. **Journal of Clinical Oncology**, American Society of Clinical Oncology, v. 17, n. 2, p. 460–460, 1999. Citado na página 2.

LOI, S. et al. Definition of clinically distinct molecular subtypes in estrogen receptor–positive breast carcinomas through genomic grade. **Journal of clinical oncology**, American Society of Clinical Oncology, v. 25, n. 10, p. 1239–1246, 2007. Citado na página 4.

MCVEIGH, T. P. et al. Assessing the impact of neoadjuvant chemotherapy on the management of the breast and axilla in breast cancer. **Clinical breast cancer**, Elsevier, v. 14, n. 1, p. 20–25, 2014. Citado na página 2.

MELCHOR, L. et al. Distinct genomic aberration patterns are found in familial breast cancer associated with different immunohistochemical subtypes. **Oncogene**, Nature Publishing Group, v. 27, n. 22, p. 3165–3175, 2008. Citado na página 4.

MELCHOR, L. et al. Estrogen receptor status could modulate the genomic pattern in familial and sporadic breast cancer. **Clinical Cancer Research**, AACR, v. 13, n. 24, p. 7305–7313, 2007. Citado na página 4.

MODLICH, O. et al. Predictors of primary breast cancers responsiveness to preoperative epirubicin/cyclophosphamide-based chemotherapy: translation of microarray data into clinically useful predictive signatures. **Journal of Translational Medicine**, BioMed Central, v. 3, n. 1, p. 1, 2005. Citado 2 vezes nas páginas 3 e 7.

NIELSEN, T. O. et al. A comparison of pam50 intrinsic subtyping with immunohistochemistry and clinical prognostic factors in tamoxifen-treated estrogen receptor–positive breast cancer. **Clinical Cancer Research**, AACR, v. 16, n. 21, p. 5222–5232, 2010. Citado na página 6.

O., T. F. **PREDITOR GÊNICO PARA A QUIMIOTERAPIA NEOADJUVANTE**. Fevereiro de 2015. 39 p. Dissertação (Mestrado em Modelagem Matemática Computacional) — Centro Federal de Educação Tecnológica de Minas Gerais, Belo Horizonte, 2015. Citado na página 8.

PAIK, S. et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. **New England Journal of Medicine**, Mass Medical Soc, v. 351, n. 27, p. 2817–2826, 2004. Citado na página 3.

PARKER, J. S. et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. **Journal of clinical oncology**, American Society of Clinical Oncology, v. 27, n. 8, p. 1160–1167, 2009. Citado 2 vezes nas páginas 4 e 5.

PEROU, C. M. et al. Molecular portraits of human breast tumours. **Nature**, Nature Publishing Group, v. 406, n. 6797, p. 747–752, 2000. Citado na página 5.

PERREARD, L. et al. Classification and risk stratification of invasive breast carcinomas using a real-time quantitative rt-pcr assay. **Breast Cancer Research**, BioMed Central, v. 8, n. 2, p. 1, 2006. Citado na página 6.

PRAT, A.; ELLIS, M. J.; PEROU, C. M. Practical implications of gene-expression-based assays for breast oncologists. **Nature reviews Clinical oncology**, Nature Publishing Group, v. 9, n. 1, p. 48–57, 2012. Citado na página 6.

PUSZTAI, L. et al. Molecular classification of breast cancer: limitations and potential. **The Oncologist**, AlphaMed Press, v. 11, n. 8, p. 868–877, 2006. Citado 2 vezes nas páginas 3 e 4.

PUSZTAI, L. et al. Individualized chemotherapy treatment for breast cancer: is it necessary? is it feasible? **Drug resistance updates**, Elsevier, v. 7, n. 6, p. 325–331, 2004. Citado na página 3.

RICHARDSON, A. L. et al. X chromosomal abnormalities in basal-like human breast cancer. **Cancer cell**, Elsevier, v. 9, n. 2, p. 121–132, 2006. Citado na página 4.

SNUSTAD, D. P.; SIMMONS, M. J. **Fundamentos da Genética**. 3. ed. Rio de Janeiro: Guanabara Koogan, 2017. v. 7. (10, v. 7). An optional note. ISBN 978-8527730860. Citado na página 2.

SØRLIE, T. et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 98, n. 19, p. 10869–10874, 2001. Citado 3 vezes nas páginas 4, 5 e 6.

SØRLIE, T. et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 100, n. 14, p. 8418–8423, 2003. Citado na página 6.

SOTIRIOU, C.; PUSZTAI, L. Gene-expression signatures in breast cancer. **New England Journal of Medicine**, Mass Medical Soc, v. 360, n. 8, p. 790–800, 2009. Citado 2 vezes nas páginas 2 e 4.

TIBSHIRANI, R. et al. Diagnosis of multiple cancer types by shrunken centroids of gene expression. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 99, n. 10, p. 6567–6572, 2002. Citado na página 6.

VINCENT-SALOMON, A. et al. Identification of typical medullary breast carcinoma as a genomic sub-group of basal-like carcinomas, a heterogeneous new molecular entity. **Breast Cancer Research**, BioMed Central, v. 9, n. 2, p. 1, 2007. Citado na página 4.