Universidade do Estado do Rio de Janeiro

Centro de Tecnologia e Ciências

Faculdade de Engenharia

Noemi da Paixão Pinto

Detecção de Alterações Respiratórias na Fibrose Cística Através da Técnica de Oscilações Forçadas e Algoritmos de Aprendizado de Máquinas

Rio de Janeiro

2018

Noemi da Paixão Pinto

Detecção de Alterações Respiratórias na Fibrose Cística Através da Técnica de Oscilações Forçadas e Algoritmos de Aprendizado de Máquinas

Dissertação apresentada, como requisito parcial para obtenção do título de Mestre, ao programa de Pós-Graduação em Engenharia Eletrônica da Universidade do Estado do Rio de Janeiro. Área de concentração: Sistemas Inteligentes e Automação.

Orientadores: Prof. Dr. Jorge Luís Machado do Amaral

Prof. Dr. Pedro Lopes de Melo

Rio de Janeiro

2018

CATALOGAÇÃO NA FONTE

UERJ / REDE SIRIUS / BIBLIOTECA CTC/B

P659 Pinto, Noemi da Paixão.
Detecção de alterações respiratórias na fibrose cística através da técnica de oscilações forçadas e algoritmos de aprendizado de máquinas / Noemi da Paixão Pinto. – 2018.
114f.

Orientadores: Jorge Luís Machado do Amaral, Pedro Lopes de Melo.
Dissertação (Mestrado) – Universidade do Estado do Rio de Janeiro, Faculdade de Engenharia.

1. Engenharia eletrônica - Teses. 2. Aprendizado do computador - Teses. 3. Teoria bayesiana de decisão estatística - Teses. 4. Algoritmos genéticos - Teses. I. Amaral, Jorge Luís Machado do. II. Melo, Pedro Lopes de. III. Universidade do Estado do Rio de Janeiro, Faculdade de Engenharia. IV. Título.

CDU 004.891

Bibliotecária: Júlia Vieira – CRB7/6022

Assinatura

Data

Noemi da Paixão Pinto

Detecção de Alterações Respiratórias na Fibrose Cística Através da Técnica de Oscilações Forçadas e Algoritmos de Aprendizado de Máquinas

Dissertação apresentada, como requisito parcial para obtenção do título de Mestre, ao programa de Pós-Graduação em Engenharia Eletrônica da Universidade do Estado do Rio de Janeiro. Área de concentração: Sistemas Inteligentes e Automação.

Aprovado em 15 de maio de 2018.

Banca Examinadora:

_____

Prof. Dr. Jorge Luís Machado do Amaral (Orientador)

Faculdade de Engenharia – UERJ

_____

Prof. Dr. Pedro Lopes de Melo (Orientador)

Instituto de Biologia – UERJ

_____

Prof. PhD. Ana Cristina Bicharra Garcia

Centro de Ciências Exatas e Tecnologia – UNIRIO

_____

Prof. Dr. Nayat Sánchez Pi

Instituto de Matemática e Estatística – UERJ

Rio de Janeiro

2018

DEDICATÓRIA

Dedico este trabalho ao Laboratório de Redes Industriais e Sistemas de Automação (LARISA) e ao Laboratório de Instrumentação Biomédica da UERJ (LIB-UERJ) pelos esforços em se manterem ativos, mesmo em meio à crise vivida no Estado do Rio de Janeiro, e por abrir espaço para o desenvolvimento de métodos para o diagnóstico e estudo de doenças que afetam o sistema respiratório, como a fibrose cística.

AGRADECIMENTOS

Agradeço a Deus por se mostrar presente em minha vida, sendo meu refúgio e fortaleza, e pelo privilégio em continuar os estudos através do mestrado. Aos meus amigos e pais, Augusto e Rose, pelo incentivo demonstrado em todas as etapas de minha vida, sempre acompanhados de muito bom humor e palavras de ânimo. À minha amiga e irmã, Laís, que me influenciou e incentivou a seguir nessa área. Por todas as palavras de ânimo que, também acompanhadas de muito bom humor, sempre trouxeram alívio nos mais diversos momentos. Ao meu amigo e companheiro de todas as horas, Sávio, que me incentivou a entrar no mestrado, me ajudou a concluir essa etapa e sempre me deu palavras de ânimo para continuar. À minha amiga, Patrícia, que conheci através do mestrado, e aos amigos Adriano e Alexandre que sempre me ajudaram por meio de explicações, palavras de ânimo e incentivo durante essa caminhada. Aos companheiros de laboratório Hugo, Everton, George e Anderson que me receberam muito bem e me fizeram sentir parte do LARISA. Pelas conversas que fizeram toda a diferença em meio às tarefas. Ao professor Pedro Lopes pela orientação, dedicação e oportunidade de continuar desenvolvendo um trabalho com aplicação na área de biomédica. Ao professor Jorge Amaral pela oportunidade, orientação, dedicação, incentivo e muita paciência demonstrados ao longo deste projeto. À FAPERJ e ao CNPq pelo apoio financeiro desse projeto.

E formou o Senhor Deus o homem do pó da terra, e soprou em suas narinas o fôlego da vida;
e o homem foi feito alma vivente.

Genesis 2:7 ACF

RESUMO

PINTO, Noemi P. *Detecção de alterações respiratórias na fibrose cística através da técnica de oscilações forçadas e algoritmos de aprendizado de máquinas*. 114f. 2018. Dissertação (Mestrado em Engenharia Eletrônica) – Faculdade de Engenharia, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2018.

Quando começou a ser estudada, a fibrose cística levava recém-nascidos a óbito em seu primeiro ano de vida. Entretanto, devido a avanços no tratamento, esses pacientes têm chegado até a fase adulta. Exames como teste de suor e espirometria, vêm sendo usados na tentativa de detectar a doença em sua fase inicial, porém esses métodos não têm sido eficientes. Sendo assim, um novo método vem sendo estudado para avaliar as propriedades mecânicas do sistema respiratório: a técnica de oscilações forçadas (FOT). A fim de comprovar a eficácia dessa nova técnica, este trabalho propõe o uso de algoritmos de aprendizado de máquinas para auxiliar a investigação e diagnóstico de alterações respiratórias na fibrose cística. Os dados fornecidos pela FOT foram aplicados nos algoritmos: K *Nearest Neighbor* (K-NN), *Radial Support Vector Machine* (RSVM), *Adaboost* (ADAB) e *Random Forest* (RF). Com o objetivo de manter uma boa acurácia e aumentar a interpretabilidade dos resultados obtidos, esses dados também foram submetidos a um algoritmo de Redes Bayesianas sintetizadas com algoritmo genético (RBGAOT). Dos experimentos realizados, a reatância respiratória fornecida pela FOT, foi o atributo que apresentou melhor desempenho individual (AUC=0,85). No experimento com oito atributos o algoritmo RBGAOT apresentou melhor desempenho (AUC=0,88). Com a aplicação dos métodos produto cruzado e seleção de variáveis, o *K*-NN e ADAB foram os algoritmos que tiveram melhores resultados (AUC=0,89). Os experimentos realizados mostraram que o uso de algoritmos de aprendizado de máquina aumentou a acurácia no diagnóstico de alterações respiratórias da fibrose cística. Já a inferência sobre as redes construídas pelo RBGAOT gerou um aumento na interpretabilidade das relações existentes entre as variáveis fornecidas pela FOT.

Palavras-chave: Fibrose cística; Técnica de oscilações forçadas; FOT; Aprendizado de máquina; Redes Bayesianas; Algoritmo genético; AUC.

ABSTRACT

PINTO, Noemi P. *Detection of respiratory changes in cystic fibrosis by forced oscillation technique and machine learning algorithms*. 114f. 2018. Dissertação (Mestrado em Engenharia Eletrônica) – Faculdade de Engenharia, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2018.

When the cystic fibrosis studies began, it used to lead newborns to death after their first year of life. However, due to advances in treatment of cystic fibrosis, these patients have reached adulthood. Medical exams such as sweat test and spirometry, have been used as an attempt to diagnose the disease on its first stage, but these methods have not been efficient. Therefore, a new method is being studied to evaluate the mechanical properties of the respiratory system: the Forced Oscillation Technique (FOT). To prove the efficiency of this new technique, the present work proposes the use of machine learning algorithms to help the investigation and diagnosis of respiratory changes in cystic fibrosis. The data provided by FOT were used on the following algorithms: K Nearest Neighbor (K-NN), Radial Support Vector Machine (RSVM), Adaboost (ADAB) and Random Forest (RF). With the purpose of keeping a good accuracy and increase the interpretability of the results, this data was submitted to Bayesian Network synthesized by genetic algorithm (RBGAOT). From the experiments performed, the respiratory reactance provided by the FOT was the feature selection that presented the best individual performance (AUC=0.85). On the experiment with eight features, the RBGAOT had the best performance (AUC=0.88). When the methods of cross product and feature selection were applied, the K-NN and ADAB were the algorithms with the best results (AUC=0.89). The experiments realized showed that the use of machine learning algorithms increased the accuracy on the diagnosis of respiratory changes in cystic fibrosis. The inference about the networks constructed by RBGAOT generated an increase in the interpretability of the existing relation between the variables provided by the FOT.

Keywords: Cystic fibrosis; Forced oscillation technique; FOT; Machine learning; Bayesian Networks; Genetic algorithm; AUC.

LISTA DE ILUSTRAÇÃO

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ADAB	Adaboost
AE	Algoritmos evolucionários
AG	Algoritmo genético
ANOVA	Análise de Variância
AUC	*Area Under the ROC curve*
BDeu	*Bayesian Dirichlet Equivalent Uniform*
CFTR	*Cystic Fibrosis Transmembrane Conductance Regulator*
DAG	*Directed Acyclic Graphs*
DPC	Distribuições de probabilidade conjunta
DPN	Diferença de potencial nasal
DPOC	Doença respiratória obstrutiva crônica
FOT	*Forced Oscillation Technique*
GAOT	*Genetic Algorithms for optimization*
K-NN	*K Nearest Neighbor*
LIB	Laboratório de Instrumentação Biomédica da UERJ
MPF	Melhor parâmetro da FOT
PGM	*Probabilistic Graphical Model*
PNT	Pneumotacômetro
RBGAOT	Redes Bayesianas sintetizadas por algoritmos genéticos
ROC	*Receiver Operating Characteristic*
RF	*Random Forest*
RSVM	*Radial Support Vector Machine*
LSVM	*Linear Support Vector Machine*
TP	Transdutor

SUMÁRIO

INTRODUÇÃO

A fibrose cística é uma doença genética que inicialmente era diagnosticada em recém-nascidos. Essas crianças eram levadas a óbito ainda no primeiro ano de vida, apresentando problemas, principalmente, no sistema respiratório. Porém, nos últimos anos houve avanço no tratamento e diagnóstico da doença, fazendo com que esses pacientes chegassem à idade adulta (LIMA et al., 2010). A espirometria é um dos principais métodos usados atualmente para o diagnóstico da fibrose cística, porém por ser um exame mais simples, não caracteriza em detalhes o sistema respiratório e não permite um melhor entendimento dos processos da doença. Sendo assim, a busca por novas técnicas tem sido uma grande motivação na pesquisa para aprimorar a identificação dessa doença.

Dentre os novos métodos pesquisados, a técnica de oscilações forçadas (FOT - *Forced Oscillation Technique*) tem sido estudada para avaliar as propriedades mecânicas do sistema respiratório (LIMA et al., 2015). A utilização dos parâmetros obtidos pela FOT, associada aos métodos de aprendizado de máquinas, trouxe importantes avanços no diagnóstico de doenças respiratórias (AMARAL et al., 2013; AMARAL et al., 2015; AMARAL et al., 2017).

Diversas publicações têm mostrado que é possível aplicar os algoritmos de aprendizado de máquinas no diagnóstico e estudo de doenças respiratórias. O artigo (AMARAL et al., 2013) descreve o uso de classificadores para aumentar a acurácia na identificação de mudanças no sistema respiratório de pacientes com tabagismo. Usando como medida de desempenho a área sob a curva ROC (AUC – *Area Under the Receiver Operating Characteristic Curve*), os algoritmos usados foram: classificadores logísticos lineares, *K*-NN, redes neurais e SVM. Dentre os testes apresentados, os melhores desempenhos foram obtidos pelo *K*-NN e SVM com valores de AUC iguais a 0,91. Esses resultados caracterizam alta taxa de acerto na classificação e comprovam que o uso de algoritmos de aprendizado de máquinas aumentou a acurácia na identificação de alterações no sistema respiratório geradas pelo tabagismo.

Já o artigo (AMARAL et al., 2015), propõe técnicas para classificar automaticamente os níveis de obstrução das vias aéreas de portadores de doença pulmonar obstrutiva crônica (DPOC). Os algoritmos *K*-NN, RF, LSVM e RSVM, foram usados durante os experimentos e avaliados de acordo com a AUC. Os classificadores *K*-NN e RF apresentaram melhor desempenho, com valores de AUC maiores que 0,9 na maioria dos procedimentos realizados. Esses resultados comprovam que o uso de algoritmos de aprendizado de máquinas pode

ajudar na categorização da obstrução das vias aéreas da DPOC e auxiliar os médicos na análise da progressão da doença.

O artigo (AMARAL et al., 2017), propõe o desenvolvimento de classificadores automáticos para simplificar o uso clínico e aumentar a precisão da FOT no diagnóstico de obstrução das vias aéreas em pacientes portadores de asma. Os algoritmos K-NN, RF, AdaBoost (ADAB) e classificador no espaço de dissimilaridade (FDSC – *Feature-based Dissimilarity Space Classifier*) foram usados e avaliados durante os experimentos, de acordo com a AUC. Os classificadores ADAB e K-NN apresentaram melhor desempenho com valores de AUC variando entre 0,88 e 0,91 durante os testes realizados. De acordo com os resultados observados, os classificadores usados podem ajudar no diagnóstico da obstrução das vias aéreas em pacientes asmáticos, auxiliando os médicos na identificação da obstrução das vias aéreas.

Embora o uso de aprendizado de máquina em associação com os parâmetros da FOT apresente elevado potencial no diagnóstico de alterações respiratórias na fibrose cística, essa associação ainda não foi investigada. Desta maneira, esse trabalho se insere na linha de pesquisa dos trabalhos citados, propondo o uso de algoritmos de aprendizado de máquinas para aprimorar ainda mais o diagnóstico e aplicação da FOT em doenças respiratórias. Dessa forma, novas informações podem auxiliar a equipe médica na investigação e diagnóstico de alterações respiratórias em portadores de fibrose cística, através dos parâmetros fornecidos pela FOT.

Atualmente, esses parâmetros são usados separadamente para realizar o diagnóstico de doenças respiratórias, sendo o atributo que apresentar maior número de acertos, o critério selecionado para identificação da doença. Nos trabalhos dessa linha de pesquisa citados anteriormente, o conjunto de dados usado para treinar os algoritmos de aprendizado de máquinas era composto por diversas amostras obtidas pela FOT, apresentando melhor desempenho do que o método atual. Apesar desses artigos comprovarem uma boa acurácia no diagnóstico de doenças respiratórias, ainda são necessários estudos que aumentem a interpretação dos resultados fornecidos por esses métodos.

Geralmente, a acurácia é a medida de desempenho usada em algoritmos de aprendizado de máquinas. Porém, há casos onde a interpretação do processo de classificação é mais, ou tão importante, quanto à previsão feita pelo modelo. Quando isso ocorre, o algoritmo escolhido precisa realizar o estudo do conjunto de dados disponível, gerando informações novas sobre o problema e facilitando o entendimento do usuário final (BRATKO, 1997). Essa característica de expressar o comportamento de um sistema de forma compreensível é

chamada de interpretabilidade e está relacionada a fatores ligados a estrutura do modelo, porém não possui uma medida padrão para ser avaliada (GACTO et al., 2011).

Mesmo que seja uma técnica de simples execução, a análise do sistema respiratório pela FOT é de difícil compreensão. Por isso é necessário treinamento e experiência da equipe médica para interpretar as curvas de resistência, reatância e os diversos parâmetros provenientes de outras medidas obtidas por essa técnica (AMARAL et al., 2013). Logo, optar por um algoritmo de aprendizado de máquinas que forneça interpretabilidade do resultado, pode agregar mais informações para o estudo e diagnóstico da fibrose cística. Sendo assim, este trabalho também propõe o uso do algoritmo de Redes Bayesianas que fornece interpretação de seus resultados por meio de grafos, realizando a tomada de decisões a partir do raciocínio baseado em probabilidades. Com estruturas gráficas, essas redes possibilitam a representação e o raciocínio sobre um domínio incerto, tornando possível lidar com a falta de informação.

A estrutura de uma Rede Bayesiana é formada por nós, que representam as variáveis do problema e são interligadas por arcos, cuja única limitação é a exigência que os grafos formados sejam acíclicos dirigidos (DAG – *Directed Acyclic Graphs*). As ligações entre as variáveis podem ser quantificadas com o cálculo das tabelas de probabilidades condicionais (SANTANA et al., 2007). Esse algoritmo tem como vantagem o fato de conseguir lidar com grande quantidade de atributos de entrada e mesmo assim apresentar essas tabelas de forma mais compacta, já que cada variável é influenciada apenas pelas variáveis diretamente ligadas a ela.

O aprendizado das Redes Bayesianas pode ser dividido em duas etapas: o aprendizado da topologia, considerado um problema complexo, e o aprendizado das probabilidades condicionais. Ambos podem ser feitos por um especialista, mas também há possibilidade de realizá-los de forma automática por meio de outros algoritmos. Nesse caso, as estruturas são geralmente construídas com base no conjunto de dados e o modelo obtido é usado para prever novos resultados (GONÇALVES, 2017).

Dentre os algoritmos usados para a busca de uma estrutura de Rede Bayesiana, podem ser destacados o *K2* e o *B* (PIFER, 2006). O algoritmo *K2* inicia sua busca com uma estrutura simples onde todas as variáveis são consideradas independentes. A cada iteração a entropia da rede é calculada e os arcos são adicionados à medida que a entropia é minimizada. As probabilidades condicionais são obtidas diretamente do conjunto de dados (HERSKOVITS et al., 1991). Já o algoritmo *B* é inicializado como o algoritmo *K2*, porém os nós e arcos são acrescentados de acordo com a diferença entre os valores de qualidade. Esse processo é

repetido até que a qualidade não aumente mais ou até que a rede esteja completa (CASTILLO et al., 1996).

A busca realizada por esses algoritmos é considerada NP-difícil devido à grande quantidade de estruturas DAG que podem descrever as relações entre suas variáveis (LARRAÑAGA et al., 1996). Motivados por essa limitação, estratégias vem sendo estudadas para a seleção de estruturas e pesquisas têm mostrado que o uso de algoritmos evolucionários (AE), fornece resultados eficientes para essa busca (TONDA et al., 2012; MYERS et al., 1999; MURUZÁBAL et al., 2007; LARRAÑAGA et al., 1996). Em geral, os AE realizam uma busca probabilística baseada nos princípios da evolução natural das espécies. Essa técnica pode ser aplicada mesmo em casos onde há muitos atributos de entrada e um conjunto de dados limitado (TONDA et al., 2012). Os AE se dividem em três principais tipos: algoritmos genéticos, programação evolutiva e estratégias de evolução (GABRIEL et al., 2008).

Este trabalho também propõe o uso de algoritmos genéticos (AG) para estimar as estruturas de Redes Bayesianas que melhor representam as relações existentes entre as características fornecidas pela FOT. Os AG são inspirados na teoria de seleção natural das espécies, proposta por Darwin, e usada para busca e otimização de problemas complexos. Dessa forma, mais informações podem ser geradas para auxiliar a equipe médica no estudo e diagnóstico de anormalidades respiratórias na fibrose cística.

Os três primeiros capítulos a seguir são destinados à revisão teórica dos principais assuntos que se baseia este trabalho. O primeiro capítulo descreve a fibrose cística, abordando suas causas, sintomas e atuais métodos de diagnóstico. O segundo capítulo apresenta os parâmetros fornecidos pela FOT e as vantagens em fazer uso dessa técnica. O terceiro capítulo descreve de forma inicial os algoritmos de aprendizado de máquinas escolhidos para aplicar os dados obtidos na FOT. O quarto capítulo apresenta o modelo proposto para este trabalho, que conta com os algoritmos descritos no capítulo 3, além da aplicação de algoritmos genéticos para gerar estruturas de Redes Bayesianas. Já o capítulo 5, mostra os resultados dos experimentos realizados com o uso das técnicas de seleção de variáveis e produto cruzado, bem como a inferência realizada nas estruturas geradas.

1. FIBROSE CÍSTICA

A mucoviscidose, ou fibrose cística, é uma doença hereditária autossômica recessiva que atinge pessoas de ambos os sexos, sendo mais incidente na população caucasiana. É causada por mutações no gene localizado no braço longo do cromossomo sete e responsável por codificar uma proteína chamada CFTR (*Cystic Fibrosis Transmembrane Conductance Regulator*) (MOTA et al., 2015; CASTELLANI et al., 2008). A CFTR é um canal responsável por regular e participar do transporte de eletrólitos por meio das membranas celulares dos sistemas respiratório, digestivo e do aparelho reprodutor, sendo o sistema respiratório o mais afetado (DALCIN et al., 2008).

Nos recém-nascidos, os primeiros sintomas podem ser observados através de alterações nas pequenas vias aéreas e tosses crônicas, logo nos primeiros meses. Já pacientes menores de 18 anos, podem apresentar quadros de pneumonia recorrentes (RIBEIRO, 2002). De uma forma geral, por se tratar de uma doença progressiva, a fibrose cística causa um aumento na obstrução do fluxo de ar, bem como o aumento da frequência respiratória e da dificuldade expiratória. Esses sintomas causam incômodo durante o sono e uma diminuição na tolerância às atividades físicas, à fisioterapia e até mesmo atividades normais realizadas no cotidiano, diminuindo assim, a expectativa de vida dos pacientes.

Quando a fibrose cística começou a ser estudada, pacientes eram levados ao óbito no primeiro ano de vida. Porém, devido ao avanço no diagnóstico e tratamento da doença, esses indivíduos têm chegado até a fase adulta. Atualmente, cerca de 70.000 pacientes estão registrados em todo o mundo. As incidências da doença variam de acordo com a região, conforme a Tabela 1, considerando pacientes da população branca de recém-nascidos (vivos) (MOTA et al., 2015).

Tabela 1 – Incidência de fibrose cística em diferentes regiões
(Extraído de: LIMA et al., 2010)

Região ou País	Incidência
Europa	1:2000 a 1:3000
Latino-Americanos	1:4000 a 1:10000
Africa do Sul	1:7056
Japão	1:350.000

Com o passar dos anos, os estudos sobre a doença foram avançando e tornaram possíveis análises mais profundas sobre o gene defeituoso, bem como a avaliação do estado do paciente, tornando possível também a expansão e avanços no tratamento. Desde então, a expectativa de vida dos portadores de fibrose cística tem se tornado cada vez maior.

Em 1938, o primeiro trabalho com a descrição da doença relatou que a expectativa de vida dos recém-nascidos era menor que um ano de idade (ANDERSEN, 1938). Já registros norte-americanos realizados em 2007 mostraram que 43% dos portadores de Fibrose Cística tinham mais de 18 anos e a idade média dos pacientes chegava a 36,5 anos de idade (DALCIN et al., 2008). Já em 2014, a expectativa média de vida dos pacientes era de 39,3 anos (LIMA et al., 2010; *Cystic Fibrosis Foundation Patient Registry*, 2014).

Atualmente, é recomendado que o diagnóstico da fibrose cística seja feito com base em três parâmetros. O primeiro parâmetro é a análise clínica que inclui: triagem neonatal, sintomas característicos da doença e histórico familiar. O segundo parâmetro é a concentração de cloreto de sódio obtido através do teste de suor, onde indivíduos com resultado menor que 30 mmol/L, são considerados indivíduos com poucas chances de portar a doença e indivíduos com resultados maiores que 60 mmol/L possuem grandes chances de portar a doença. Já indivíduos com concentração de cloreto na faixa de 30 a 59 mmol/L, devem passar pela análise do terceiro parâmetro: a avaliação da proteína CFTR. No caso da descoberta de duas ou mais mutações nessa proteína, há fortes indícios de se tratar de um portador da doença. Se for identificada apenas uma mutação, é feita uma análise mais profunda da disfunção da CFRT através da medição da diferença de potencial nasal (DPN). O fluxograma da Figura 1 mostra um resumo das recomendações para o diagnóstico da fibrose cística (FARRELL et al., 2017).

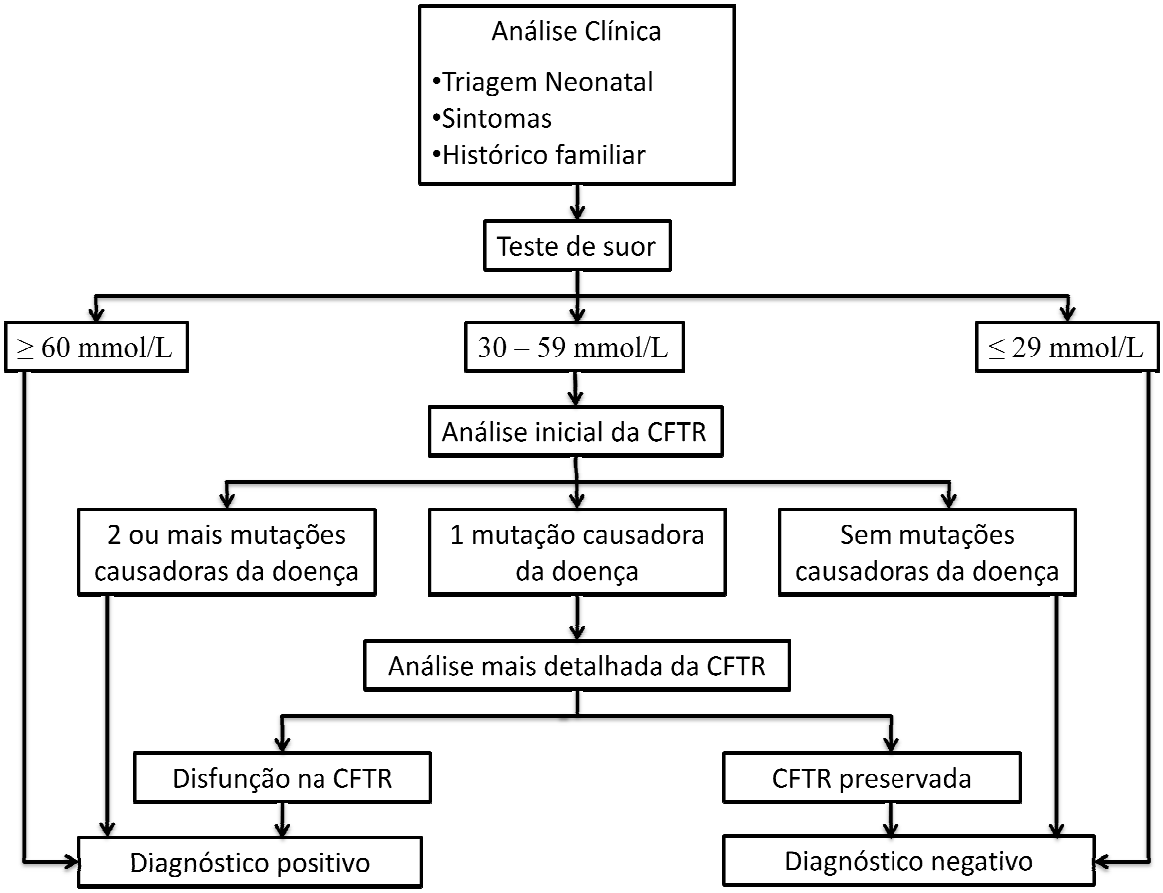

Figura 1 – Fluxograma de recomendações para o diagnóstico da fibrose cística
(Adaptado de FARRELL et al., 2017)

Outra ferramenta usada para complementar a investigação e diagnóstico de alterações respiratórias na fibrose cística é a espirometria, avaliando as alterações respiratórias dos pacientes por meio de fluxos e volumes respiratórios. Dentre os resultados obtidos, especificamente o volume expiratório forçado no primeiro segundo, é o mais usado na tentativa de prever e identificar sintomas característicos em pacientes de diferentes idades. Porém, a espirometria não tem sido suficiente para diagnosticar alterações respiratórias em portadores dessa doença e uma nova técnica, que avalia as propriedades resistivas e reativas do sistema respiratório, vem sendo estudada: a técnica de oscilações forçadas (LIMA et al., 2010).

2. TÉCNICA DE OSCILAÇÕES FORÇADAS

Com o intuito de desenvolver uma análise mecânica do sistema respiratório de forma mais simples, mas que ainda apresentasse novas informações, no ano de 1956 foi proposto o uso da técnica de oscilações forçadas (FOT – *Forced Oscillation Technique*) (DUBOIS et al., 1956). Durante um ensaio da FOT, o paciente deve permanecer sentado, fazer o uso de um *clip* nasal e respirar de forma espontânea, enquanto um fluxo constante (*bias flow*) renova o ar inspirado pelo mesmo. Pequenas oscilações de pressão são geradas por um aparelho externo, normalmente um alto falante, e aplicadas às vias aéreas do paciente, que permanece respirando espontaneamente. Essa pressão *P* é medida por um transdutor (TP) e estimula um fluxo oscilatório (*V'*), medido por um pneumotacômetro (PNT). O valor da resistência total do sistema respiratório, chamada impedância de entrada (Z_{rs}), é calculado pelos sinais obtidos por TP e PNT (Figura 2) (MELO et al., 2000).

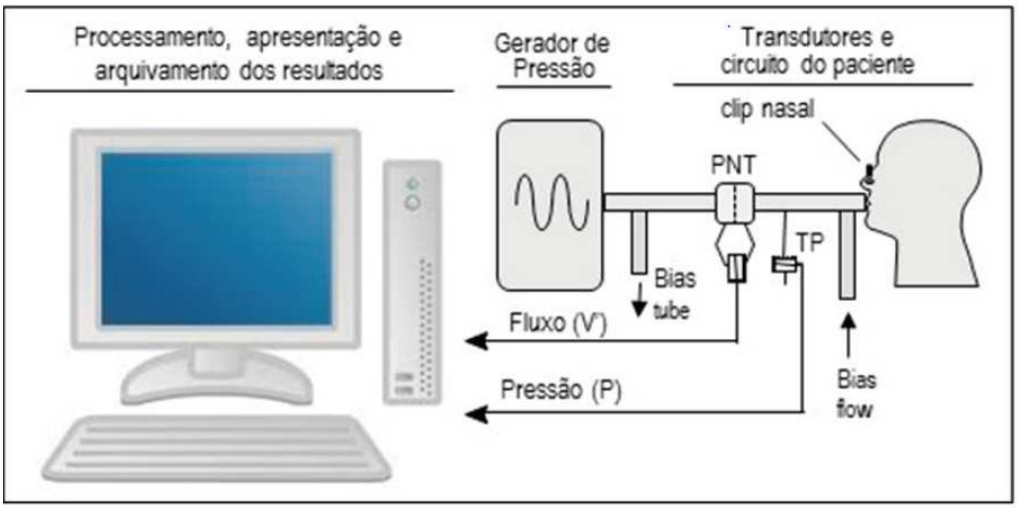

Figura 2 – Diagrama em blocos básico do sistema
(MELO, 2015)

Para diminuir o tempo de execução dos ensaios da FOT, um sistema responsável por analisar impedâncias realiza o processamento dos sinais obtidos pelas oscilações nas faixas de frequências desejadas. Pela transformada de Fourier, é possível decompor os sinais *P* e *V'* e também realizar uma avaliação das alterações do módulo de Z_{rs} em diversas frequências, conforme equação (1) (LIMA et al., 2010).

$$Zrs\ (f) = \frac{FFT\ (P)}{FFT\ (V')} \tag{1}$$

Sendo,
FFT(*P*): Transformada de Fourier da pressão *P*
FFT(*V'*): Transformada de Fourier do fluxo *V'*
f: frequência desejada

As funções senoidais provenientes da decomposição dos sinais *P* e *V'* podem ser representadas com suas respectivas componentes, conforme as equações (2) e (3):

$$P = P_m sen(\omega t) \tag{2}$$

$$V' = V'_m sen(\omega t + \varphi) \tag{3}$$

Sendo:
P_m : amplitude do sinal *P*
V'_m: amplitude do sinal *V'*
ω: frequência angular igual a 2πf
φ: diferença de fase entre os sinais *P* e *V'*

A variável Z_{rs} representa toda a carga mecânica, que inclui os efeitos das propriedades resistivas, elásticas e inertivas do sistema respiratório. Normalmente durante um ensaio da FOT, as impedâncias Z_{rs} são representadas por componentes reais e imaginários, descritos respectivamente pela resistência respiratória (R_{rs}) e pela reatância respiratória (X_{rs}) (MELO, 2015), conforme equação (4):

$$Z_{rs} = \sqrt{{R_{rs}}^2 + {X_{rs}}^2} \tag{4}$$

Os componentes de R_{rs} e X_{rs} podem ser derivados da seguinte forma (MACLEOD et al., 2001):

$$R_{rs} = |Zrs| \cos\varphi \quad (5)$$

$$X_{rs} = |Zrs| \sin\varphi \quad (6)$$

A energia cinética usada durante a aceleração do fluxo aéreo é descrita por meio da inertância respiratória (I_{rs}). Essa variável é normalmente desprezada em análises realizadas em baixas frequências, sendo o sistema respiratório modelado apenas por um componente resistivo e um complacente. Já em frequências mais elevadas, como ocorre na FOT, a I_{rs} torna-se relevante, permitindo a obtenção de informações mais detalhadas sobre as características mecânicas do aparelho respiratório com base na reatância (MELO, 2015).

A faixa de frequência de 4 a 32Hz, é a mais utilizada durante os ensaios da FOT. Nesse intervalo, a resistência respiratória caracteriza a dissipação total da energia do sistema, que abrange a soma dos efeitos vindos de resistências relacionadas a quatro fatores: ao tecido pulmonar, à parede torácica, às vias aéreas e à redistribuição do fluxo do gás nos pulmões. A reatância respiratória (X_{rs}) caracteriza o armazenamento de energia potencial do sistema que está associada à complacência respiratória (C_{rs}), sendo o armazenamento de energia cinética associado à inertância I_{rs}. As propriedades elásticas estão associadas à complacência (C_{din}) e à elastância dinâmica (E_{din}), sendo o parâmetro E_{din}, o inverso de C_{din} (equação (7)) (MELO et al., 2000). A relação entre X_{rs}, I_{rs} e C_{rs}, é descrita na equação (8):

$$E_{din} = \frac{1}{C_{din}} \quad (7)$$

$$X_{rs} = \omega I_{rs} - j\frac{1}{\omega C_{rs}} \quad (8)$$

Sendo, $\omega=2\pi f$ e $j=\sqrt{-1}$.

Devido à complexidade do sistema respiratório, é comum que as componentes de Z_{rs} não estejam na mesma fase. Porém na frequência de ressonância (F_r), os efeitos da complacência e inertância são iguais e, consequentemente, X_{rs} é igual a zero (MIRANDA et al., 2013).

É possível analisar a resistência e a reatância de forma mais detalhada usando diversas faixas de frequências. Mesmo sendo um método mais lento ainda é vantajoso, pois os valores obtidos representam as médias dos resultados do sistema respiratório durante vários períodos de ventilação. Como não há consenso na literatura sobre as características a serem avaliadas nesses valores médios, há grupos que realizam a análise pelo método de regressão linear para uma faixa de frequências de 4 a 16Hz. Dessa forma, é possível determinar a resistência no intercepto em 0Hz (R_o) e o coeficiente angular da curva de resistência (S) (LIMA et al., 2015; AMARAL et al., 2017).

O parâmetro R_o estima como as resistências newtonianas associadas às vias aéreas e aos tecidos, bem como sua resistência tardia proveniente da distribuição do gás, reagem em frequências baixas. Já o parâmetro S está associado à alteração na distribuição do fluxo de gás dentro do sistema respiratório de acordo com a frequência utilizada (MIRANDA et al., 2013). A Tabela 2 mostra um resumo de todas as características fornecidas pela FOT.

Tabela 2 – Parâmetros fornecidos pela FOT

Parâmetro	Descrição do Parâmetro
R_o	Resistência no Intercepto
R_m	Resistência Média
X_m	Reatância Média
C_{din}	Complacência Dinâmica
S	Inclinação da Curva de Resistência
Z_{4Hz}	Impedância em 4Hz
F_r	Frequência de Ressonância
E_{din}	Elastância Dinâmica

Em suma, a FOT possui duas principais vantagens. Primeiramente, esse método permite uma análise mais detalhada do sistema respiratório, fornecendo parâmetros que não podem ser obtidos pela espirometria e demais técnicas usadas. Sendo assim, a FOT apresenta forte potencial para diagnósticos e contribui para melhor compreensão dos processos da doença. Outra grande vantagem desse método é a fácil execução do exame para o profissional, dependendo apenas da cooperação do paciente que deve respirar de forma espontânea, conforme mostrado na Figura 3.

Por ser uma técnica nova que investiga as alterações mecânicas no sistema respiratório, são necessários os mais diversos estudos e testes para que se comprove sua eficácia. Assim sendo, diversas pesquisas têm sido feitas para mostrar que é possível aplicar a FOT para auxílio da equipe médica (RIBEIRO et al., 2018; MARINHO et al., 2017; LACERDA et al., 2017).

Figura 3 – Indivíduo realizando ensaios pela técnica de oscilações forçadas
(MELO et al., 2015)

3. ALGORITMOS DE APRENDIZADO DE MÁQUINAS

Com o intuito de auxiliar a equipe médica no diagnóstico de alterações respiratórias na fibrose cística por meio da FOT, foi usada a técnica de aprendizado de máquinas. Cinco algoritmos foram escolhidos para realização dos testes: *K-Nearest Neighbor*, *Random Forest*, Adaboost, *Support Vector Machine* e Redes Bayesianas. Cada um desses algoritmos foi descrito neste capítulo, com suas principais vantagens e desvantagens.

3.1. *K-Nearest Neighbor*

O algoritmo dos *K* vizinhos mais próximos (*K*-NN – *K Nearest Neighbor*) é considerado um dos mais simples algoritmos de aprendizado de máquinas. Esse algoritmo tem o aprendizado por instância, onde o conjunto de treinamento é armazenado durante o estágio de aprendizado. Quando uma nova instância precisa ser classificada, o algoritmo encontra as *K* instâncias de treinamento mais próximas, usando uma função de similaridade, que normalmente é a distância euclidiana. Uma instância *x* pode ser representada como um vetor de atributos c_r, conforme a equação (9) (FACELI et al., 2011):

$$c_r\,(x) = (c_1(x), c_2(x), c_3(x), \ldots\ c_n(x)\,) \tag{9}$$

Sendo $c_n(x)$ o valor de cada atributo do vetor $c_r(x)$.

O método *K*-NN assume que todas as instâncias estão dentro de um espaço *n*-dimensional $\Re^n$. Dessa forma, o cálculo da distância euclidiana entre duas instâncias x_i e x_j é feito conforme a equação (10):

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{n} \left(c_r(x_i) - c_r(x_j)\right)^2} \tag{10}$$

Durante o treinamento da configuração mais simples, considerando apenas 1 vizinho mais próximo (1-NN), o algoritmo aprende um conjunto de dados (*D*) com seus respectivos rótulos. Para atribuir uma classe a uma amostra de teste (*a*) não rotulada, é feito o cálculo da

distância entre o vetor de características de *a* e o vetor de características de todas as instâncias do conjunto *D*, armazenadas pelo algoritmo. O vetor de característica com menor distância classifica a amostra de teste. A Tabela 3 mostra um algoritmo básico da configuração 1-NN (SMOLA et al., 2008). A Figura 4 mostra um exemplo da configuração 1-NN aplicada a um problema de classificação, onde uma amostra pode receber o rótulo positivo ou negativo. O asterisco representa a amostra a ser classificada e, nesse exemplo, devido à proximidade com um ponto da classe negativa, a amostra em questão é classificada como negativa.

Tabela 3 – Algoritmo básico da configuração 1-NN
(Adaptado de: FACELI et al., 2011)

Conjunto de dados para treinamento: $D = \{(x_i,y_i)\}$, i = 1... n
Uma amostra que se deseja classificar: $a = \{x_a, y_a=?\}$
Distância entre as instâncias: $d(x_i,x_j)$
Resultado da classificação da amostra a: y_a
$d_{mín} \leftarrow +$ □
Para i=1:n faça
Se $d(x_i,x_j) < d_{mín}$
$d_{mín} \leftarrow d(x_i,x_j)$
ind $\leftarrow i$
Fim
Fim
$y_a = y_{ind}$

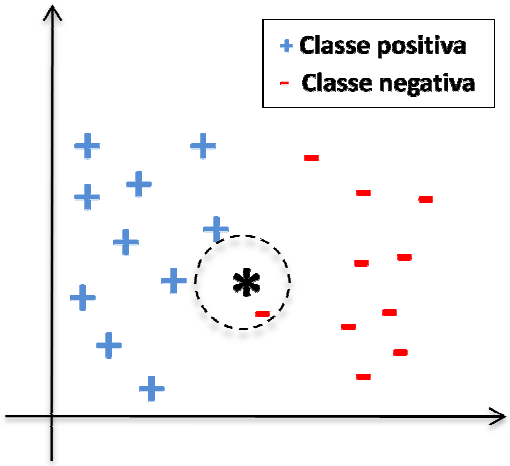

Figura 4 – Exemplo da configuração 1-NN
(Adaptado de: FACELI et al., 2011)

Na Figura 5, um exemplo das configurações 3-NN e 5-NN mostram que dependendo do número de vizinhos considerados, a mesma amostra poderia ser classificada com diferentes rótulos. Na configuração 3-NN (Figura 5 (a)), a amostra é classificada como positiva, pois está próxima a dois pontos dessa classe, mas apenas a 1 ponto da classe negativa. Já na configuração 5-NN (Figura 5 (b)), a classificação seria negativa devido à proximidade da amostra com três pontos dessa classe. A única restrição para a escolha do *K* é que seja um valor ímpar, para realizar a classificação de uma amostra sem que haja empate entre as classes.

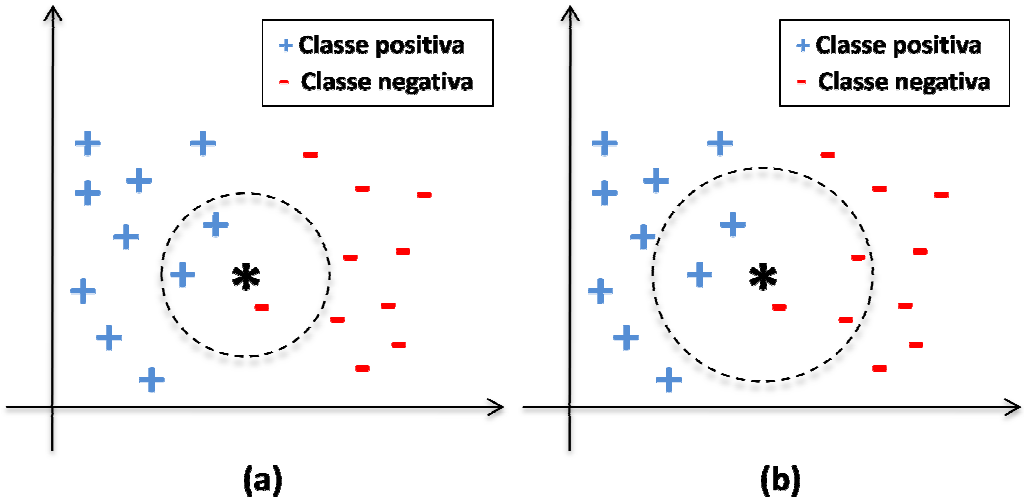

Figura 5 – Exemplo das configurações: (a) 3-NN e (b) 5-NN
(Adaptado de: FACELI et al., 2011)

Dentre as vantagens do algoritmo *K*-NN, pode-se destacar seu treinamento simples que se resume ao armazenamento dos dados de consulta em sua memória. Outra vantagem é sua aplicabilidade em problemas mais complexos, já que realiza aproximações para cada nova instância a ser classificada. Como desvantagem, pode ser citado o custo computacional em grandes conjuntos de treinamento, pois é necessário calcular a distância entre a amostra a ser classificada e cada um desses pontos. Outra desvantagem é a sensibilidade quanto à quantidade de atributos usados no problema, já que o número de atributos define a dimensão do espaço $\Re^n$.

3.2. *Random Forests*

As florestas aleatórias (*Random Forests*) são comitês de árvores de decisão e se baseiam em dois conceitos principais (BREIMAN, 2001). O primeiro é a seleção aleatória dos atributos de entrada para a formação de diversos subconjuntos, que serão submetidos às árvores de decisões. Esse processo contribui para a redução da correlação entre as diversas árvores construídas (COSTA, 2012). O segundo conceito é o *bagging* (*Bootstrap Aggregation*), usado para criação desses subconjuntos através da amostragem por *bootstrap*, onde o mesmo número de amostras do conjunto original é selecionado com repetição para cada novo subconjunto (LIAW et al., 2002). Dessa forma, podem existir tanto amostras repetidas, quanto amostras não inclusas durante o treinamento. O resultado das diversas árvores criadas é combinado, reduzindo então, a variância do resultado final fornecido pelo algoritmo.

O funcionamento do algoritmo *Random Forest* (Tabela 4) tem início com a seleção aleatória de um subconjunto *Z*, formado por amostras dos dados de treinamento com o total de *p* atributos. Em seguida, uma árvore T_b é construída em três etapas: seleção aleatória de *m* dos *p* atributos ($m << p$), escolha do melhor ponto de corte dentre os atributos selecionados e divisão de um nó em dois nós filhos, com base nesse ponto de corte. Esse procedimento é repetido para cada novo nó até alcançar o tamanho mínimo de nós ($n_{mín}$). Com as árvores de decisão construídas, é possível configurar o algoritmo para regressão ou classificação, nesse caso, a previsão final dada pelo algoritmo será a previsão dada pela maioria das árvores individuais (HASTIE et al., 2008).

Como vantagem, o algoritmo *Random Forest* consegue lidar com um grande número de variáveis de entrada mantendo sua rapidez na construção das redes. Como a escolha dos atributos de entrada é aleatória, as árvores construídas são descorrelacionadas e, consequentemente, outra vantagem é uma diminuição na variância da combinação das árvores. As desvantagens do método estão na sensibilidade a muitos ruídos na base de dados e na dificuldade de interpretação do modelo (COSTA, 2012).

Tabela 4 – Algoritmo básico do *Random Forest*
(HASTIE et al., 2008)

Para b = 1, ..., *B* faça

Seleciona um subconjunto *Z* com dados de treinamento

Constrói uma árvore de decisão seguindo as três etapas:

1. Seleciona *m* atributos
2. Define o melhor atributo dentre *m* para ponto de corte
3. Divide o nó em dois nós filhos

Fim

Saída: Conjunto de árvores $\{T_b\}$

- Para classificação: Sendo $\widehat{\boldsymbol{C}}_{\boldsymbol{b}}(\boldsymbol{x})$ a classe de um ponto $\boldsymbol{x}$ a ser classificado, tem-se:

$$\widehat{\boldsymbol{C}}_{\boldsymbol{b}}(\boldsymbol{x}) = \boldsymbol{maioria\ dos\ votos}\left\{\widehat{\boldsymbol{C}}_{\boldsymbol{b}}(\boldsymbol{x})\right\}_{\boldsymbol{1}}^{\boldsymbol{B}}$$

- Para regressão: $\hat{\boldsymbol{f}}_{\boldsymbol{RF}}^{\boldsymbol{B}}(\boldsymbol{x}) = \frac{\boldsymbol{1}}{\boldsymbol{B}}\sum_{\boldsymbol{b=1}}^{\boldsymbol{B}} \boldsymbol{T}_{\boldsymbol{b}}(\boldsymbol{x})$

3.3. Adaboost

Adaptative Boosting (Adaboost) é uma técnica de aprendizado de máquina cujo objetivo é criar um classificador forte (alta acurácia), através da combinação de vários classificadores simples (baixa acurácia). Esses classificadores são treinados em sequência e a cada novo modelo os ajustes são feitos aumentando a probabilidade dos pontos classificados de forma errada, aparecerem no próximo conjunto de treinamento (MARGINEANTU et al., 1997).

Essa probabilidade é calculada da seguinte forma: em um conjunto de treinamento $D=\{x_i,y_i\}$, x_i representa o vetor com os dados de entrada no sistema e y_i representa seus respectivos rótulos $\in \{-1,+1\}$. A cada iteração (t), é calculada uma distribuição (D_t):

$$D_{t+1}(i) = \frac{D_t(i)\, exp\left(-\alpha_t y_i h_t(x_i)\right)}{\sum_{i=1}^{n} D_t(i)} \tag{11}$$

Com a distribuição calculada, um algoritmo simples é aplicado para encontrar uma hipótese (h_t). Pelo erro ε_t, o peso do classificador simples (α_t) pode ser calculado através da equação (12), onde as hipóteses com menor valor de erro recebem maior peso α_t.

$$\alpha_t = \frac{1}{2}\ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right) \tag{12}$$

Em seguida, obtém-se o resultado final $H(x)$ pela aplicação da função sinal na combinação ponderada de todas as hipóteses h_t (equação (13)). O algoritmo básico que descreve o funcionamento do Adaboost está na Tabela 5 (SCHAPIRE, 2013).

$$F(x) = \sum_{t=1}^{T} \alpha_t h_t(x) \;\rightarrow\; H(x) = sign\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right) \tag{13}$$

Tabela 5 – Algoritmo básico do Adaboost
(Adaptado de: SCHAPIRE, 2013)

Conjunto de dados para treinamento: $D = \{(x_i, y_i)\}$
Inicializa: $\boldsymbol{D_1(i) = \frac{1}{m}}$ para i=1, ..., m
Para t = 1, ..., T faça
Treina o classificador através de D_t
Calcula as hipóteses h_t
Seleciona h_t com menor erro $\boldsymbol{\varepsilon_t}$
Escolha do α_t
Atualiza o valor de D_t para i=1, ..., T
Fim
Saída $H(x)$

O Adaboost possui como vantagem sua simples implementação, já que o algoritmo utiliza classificadores simples e estes, sucessivamente, vão se especializando em acertar a classificação que os classificadores anteriores fizeram de forma errônea. A única restrição a ser feita é que os classificadores simples devem ter um desempenho superior aos classificadores aleatórios. Outra vantagem é sua boa generalização, sendo adequada para qualquer problema de classificação. Dentre as desvantagens do Adaboost estão o risco de *overfitting* durante o treinamento e sua sensibilidade ao lidar com dados ruidosos.

3.4. *Support Vector Machines*

O algoritmo *Support Vector Machines* (SVM) é baseado na teoria de aprendizagem estatística no qual, são aplicados princípios matemáticos para auxiliar a seleção de um classificador específico (*h*), por meio de seu desempenho e complexidade, a partir de um conjunto de treinamento (*D*) (FACELI et al., 2011). O SVM tem como ideia principal a criação de um hiperplano como uma fronteira de classificação, onde a margem, definida como a distância entre um ponto *x* e essa fronteira, é maximizada (Figura 6). Os limiares dessa fronteira são conhecidos como vetores de suporte (KUNCHEVA, 2014).

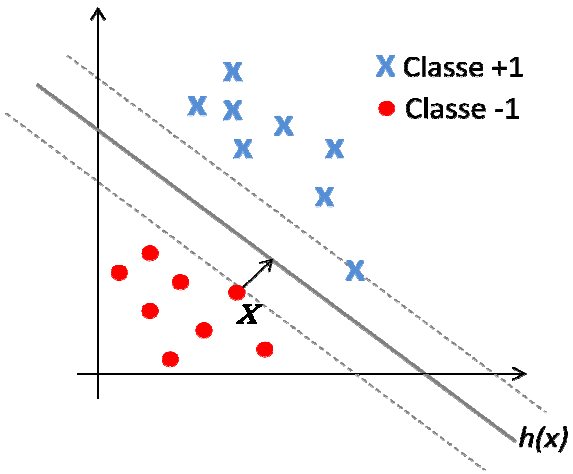

Figura 6 – Exemplo de fronteira de decisão e os vetores de suporte do algoritmo SVM
(Adaptado de: KUNCHEVA, 2014)

Quanto maior a margem selecionada, melhor será a capacidade de generalização do SVM (KUNCHEVA, 2014). Considerando um conjunto de treinamento linearmente separável $D = \{x_i, y_i\}$, onde x_i são as entradas e y_i os rótulos das classes representados por dois valores possíveis: -1 ou 1, a fronteira de classificação é dada por um hiperplano representado pela equação (14):

$$□(x) = w.x + b \tag{14}$$

Onde:

w: vetor normal ao hiperplano

b: número escalar

w.x: produto escalar

A equação (14) divide o espaço dos dados de entrada em duas regiões:

$$w.x + b > 0 \quad e \quad w.x + b < 0$$

Para classificar um ponto *x*, é necessário aplicar uma função sinal (*sgn*) em *h(x)*:

$$g(x) = sgn(h(x)) = \begin{cases} +1, w.x + b > 0 \\ -1, w.x + b < 0 \end{cases} \tag{15}$$

Ao multiplicar o vetor *w* e a constante *b* na equação (14) por uma mesma variável, é possível obter diversos hiperplanos correspondentes. Sendo assim, *w* e *b* geram hiperplanos onde: $h(x) = w.x+b \geq 0$, quando y_i=+1 e $h(x) = w.x+b < 0$, quando y_i=-1. Dessa forma, podem-se escrever as equações no sistema (16) e observar sua ilustração da Figura 7 (SMOLA et al., 2008):

$$\begin{cases} w.x_i + b \geq +1 \; se \; y_i = +1 \\ w.x_i + b \leq -1 \; se \; y_i = -1 \end{cases} \tag{16}$$

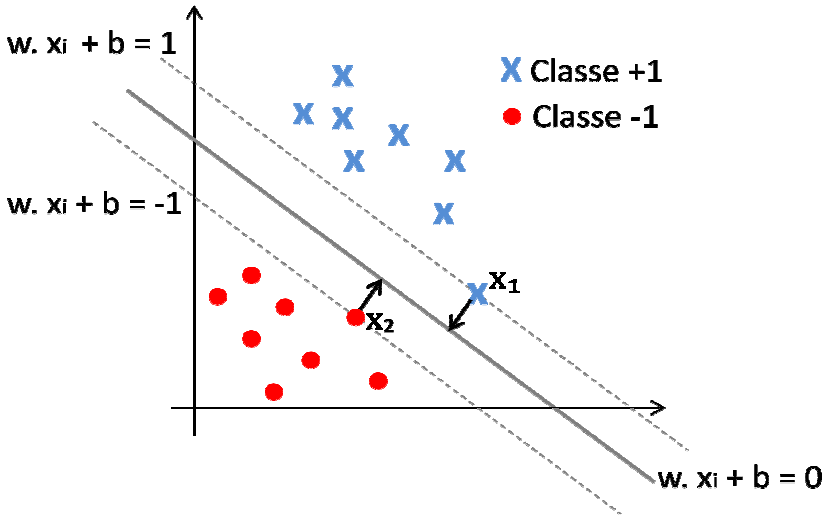

Figura 7 – Exemplo de hiperplanos na classificação SVM
(Adaptado de: FACELI et al., 2011)

Para realizar o cálculo da margem na Figura 7, é necessário subtrair a equação do hiperplano de x_2 da equação do hiperplano de x_1:

$$\begin{cases} w.x_1 + b = +1 \\ w.x_2 + b = -1 \end{cases} \rightarrow w(x_1 - x_2) = 2 \rightarrow \|x_1 - x_2\| = \frac{2}{\|w\|} \quad (17)$$

Para maximizar a margem devem-se minimizar os pesos. A fim de facilitar esse cálculo foram feitas alterações matemáticas para que um problema de maximização fosse reduzido a um problema de minimização de uma função quadrática, que representa a função de custo:

$$min\|w\| \rightarrow min\ \frac{1}{2}\|w\|^2 \quad (18)$$

O algoritmo SVM linear também é chamado de SVM com margens rígidas, pois são estabelecidas restrições para certificar que pontos do conjunto de treinamento não estejam entre as margens que separam as classes do problema. Porém, para aplicações em problemas reais, dificilmente são encontrados dados linearmente separáveis. Nesses casos, é possível realizar a adição de variáveis de folga (ξ_i) que suavizam as restrições lineares, permitindo que alguns pontos de treinamento estejam entre os hiperplanos (Figura 8) e também permitem erros na classificação, como ruídos.

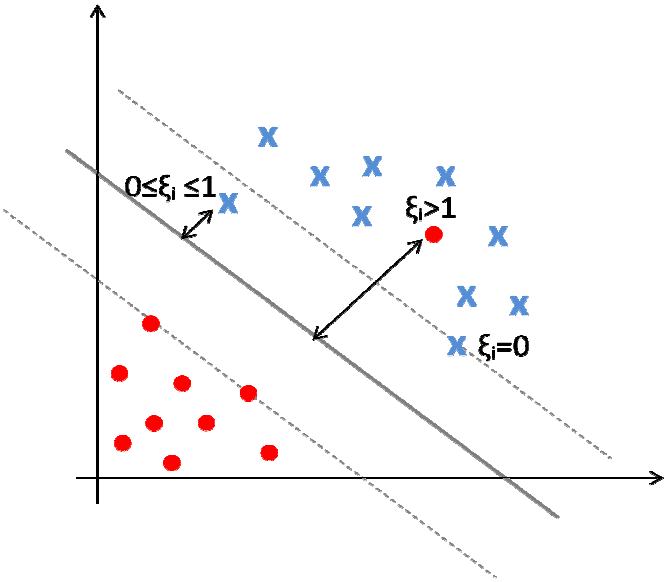

Figura 8 – Exemplo de hiperplanos na classificação SVM com margens suaves
(Adaptado de: KUNCHEVA, 2014)

Dessa forma, há diferentes situações que podem ser observadas na classificação de uma amostra. Quando $\xi_i > 1$, a amostra está fora da região de separação e do lado incorreto de sua classificação. Quando $0 < \xi_i \leq 1$, a amostra está classificada corretamente, mas entre as margens de separação. Quando $\xi_i = 0$, a amostra está sobre as margens de separação (FACELLI et al., 2011). As equações do sistema (16) podem ser reescritas conforme o sistema (19). Por flexibilizar restrições, essa configuração é chamada de SVM com margens suaves.

$$\begin{cases} w.x_i + b \geq +1 - \xi_i \; se \; y_i = +1 \\ w.x_i + b \leq -1 + \xi_i \; se \; y_i = -1 \end{cases} \tag{19}$$

Existem problemas cujos dados não podem ser divididos por um hiperplano. Nesses casos, realiza-se o mapeamento do conjunto de treinamento para um espaço de dimensão maior, chamado espaço de características. Nesse novo espaço, espera-se que esses dados sejam linearmente separáveis (Figura 9). Após o mapeamento, é usada a configuração do SVM linear com margens suaves para lidar com classificações erradas ou ruídos (HASTIE et al., 2008).

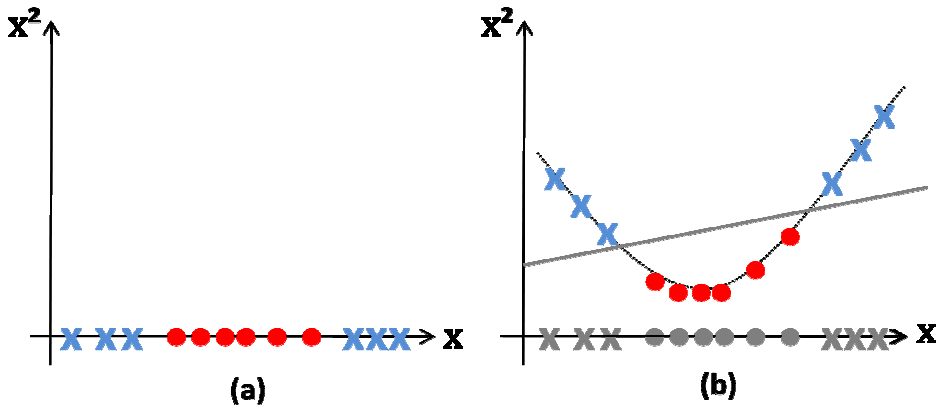

Figura 9 – Conjunto de dados: (a) em um espaço unidimensional e não linearmente separável e (b) em um novo espaço bidimensional e linearmente separável
(Adaptado de: KUNCHEVA, 2014)

Esse mapeamento em dimensões mais altas é feito através de funções de *Kernels*. O uso dessas funções na representação de espaços de características é realizado quando não se tem o conhecimento do mapeamento a ser feito. Há três principais tipos de funções *Kernel*: Polinomial, Gaussiana e *Radial Basis Function* (RBF) (FACELI et al., 2011; KUNCHEVA, 2014).

Uma das vantagens do algoritmo SVM é a sua eficiência em encontrar a melhor solução possível, já que sua função objetivo é convexa e possui apenas um mínimo global. Outra vantagem é poder aplicar o SVM em problemas que possuem elevado número de atributos, bem como em problemas de regressão. Como desvantagens, o SVM não permite uma interpretação da decisão tomada pelo algoritmo e também se mostra sensível quanto à escolha de seus parâmetros. O conjunto de treinamento do SVM precisa conter apenas dados numéricos, sendo necessário converter atributos discretos.

3.5. Redes Bayesianas

O nome destas redes é derivado da regra de Bayes, estabelecida por Thomas Bayes, que mostra como um efeito *E* transforma a probabilidade à priori P(*Cj*) em uma probabilidade à posteriori P(*Cj*|*E*) (equação (20)), alterando assim, a estimativa inicial *Cj* com base na nova informação fornecida por *E*.

$$P(C_j|E) = \frac{P(E|C_j)P(C_j)}{P(E)} \tag{20}$$

O termo P(*E*|*Cj*) é a probabilidade condicional do efeito *E* ser observado, dado a causa *Cj*. P(*E*) é um fator de normalização dado por um somatório, como mostrado na equação (21), e pode ser desprezado (SILVA et al., 2016). Os cálculos dessas probabilidades representam as relações causais entre as variáveis do problema, permitindo um aumento na interpretabilidade (SANTANA et al., 2007).

$$P(E) = \sum_{j=1}^{n} \mathrm{P}(\mathrm{E}|C_j)\mathrm{P}(C_j) \tag{21}$$

A representação em uma Rede Bayesiana mostra de forma simplificada as relações de causalidade entre as variáveis de um sistema. Essa representação é feita por nós, que correspondem às variáveis do problema, e por arcos que correspondem às conexões entre essas variáveis (Figura 10), mostrando a dependência direta entre elas (MARQUES et al., 2003).

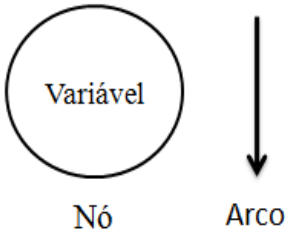

Figura 10 – Elementos de representação das Redes Bayesianas

Na nomenclatura usada para os elementos das Redes Bayesianas, alguns termos são comuns para indicar a hierarquia dos nós da rede. Os termos pai e filho mostram a dependência direta entre dois ou mais nós por meio de um arco. O nó de onde parte o arco é chamado nó pai. O nó em que o arco chega, é chamado nó filho. Na rede da Figura 11, *A* é dito pai do nó *B*. Por sua vez, o nó *B* é dito filho do nó *A*. Da mesma forma, o nó *B* é chamado pai de *C* e *D*, e ambos são filhos do mesmo nó *B*. Os nós que não estão diretamente ligados pelos arcos podem ser chamados de nós antecedentes ou nós descendentes. Usando também como exemplo a rede da Figura 11, *A* é dito antecedente de *C* e *D*, bem como *C* e *D* são descendentes de *A* (ARA-SOUZA, 2010).

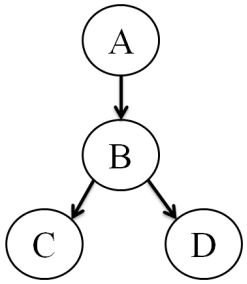

Figura 11 – Topologia de uma Rede Bayesiana simples

Outros termos usados são: nó raiz, nó folha e nó intermediário. Os nós raízes representam a origem do problema e não possuem pais. Os nós folhas mostram o resultado final do problema e não possuem filhos. Os nós que não são raízes e nem folha, são chamados de nós intermediários.

Há uma propriedade relativa a nós pais, filhos e descendentes, chamada propriedade de Markov, que diz: “não existem dependências diretas no sistema que está sendo modelado, que não sejam explicitamente mostradas nos arcos”, ou seja, uma variável é condicionalmente independente de todos os outros nós da rede que não sejam seus antecedentes (NEAPOLITAN, 2003). Atender essa propriedade é importante no uso das Redes Bayesianas, pois elas simplificam o cálculo das relações existentes entre as variáveis do problema, considerando apenas os nós que exercem influência sobre seus descendentes (KORB et al., 2011).

3.5.1. Distribuição de Probabilidade Conjunta

As relações entre os nós podem ser quantificadas através do cálculo da probabilidade condicional. É necessário olhar para cada um dos nós, ou variáveis do sistema, e analisar todas as possíveis combinações de valores dos nós pais em relação aos seus nós filhos. Vale ressaltar que a soma das probabilidades deve somar 1 para cada um dos possíveis estados de uma variável. O cálculo dessas probabilidades dá origem à tabela de distribuição de probabilidade conjunta (MARQUES et al., 2003).

Se uma Rede Bayesiana satisfaz a propriedade de Markov, em que cada nó depende apenas dos seus nós pais, então o cálculo da distribuição de probabilidade pode ser escrito como na equação (22):

$$P(x_1, x_2, \dots, x_n) = \prod_i P(x_i \,|\mathrm{Pais(X_i)}) \tag{22}$$

Uma tabela de distribuição de probabilidade conjunta representa a descrição completa do domínio de um sistema. Sendo assim, até os nós raízes possuem uma tabela que represente suas probabilidades à priori.

3.5.2. Tipos de Inferência Bayesiana

Para o raciocínio em problemas através de métodos probabilísticos, é necessário o uso do cálculo da probabilidade à posteriori sobre uma variável consulta (*query*), dada uma evidência forte (*Hard Evidence* ou *Evidence*). Dependendo do objetivo desejado, há pelo menos três formas de inferir sobre uma Rede Bayesiana e calcular a probabilidade P(*Query*|*Evidence*).

Quando o objetivo é descobrir as causas do problema, o raciocínio é feito da causa em direção ao efeito (Figura 12 (a)), seguindo a mesma direção dos arcos da rede. Quando o objetivo é o diagnóstico, o raciocínio é feito a partir dos efeitos em direção a causa (Figura 12(b)). Esse raciocínio segue o sentido oposto ao dos arcos da estrutura da rede. Se o objetivo for descobrir as causas de um efeito em comum, deve ser usado o raciocínio intercausal, onde são analisadas as variáveis *query* (*Q*) e *hard evidence* (*E*) (Figura 12 (c)) (MARQUES et al., 2003).

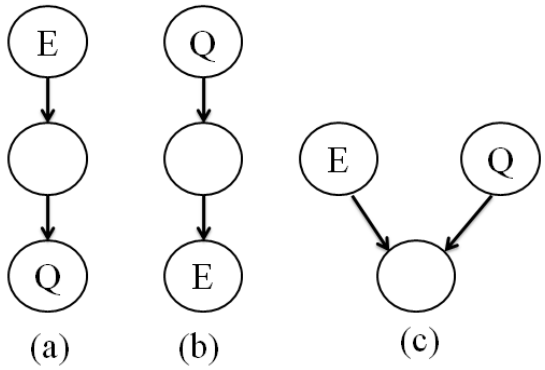

Figura 12 – Tipos de inferências bayesianas: (a) Causal; (b) Diagnóstico; (c) Intercausal;
(Adaptado de: MARQUES et al., 2003)

O exemplo a seguir mostra como realizar a inferência diagnóstica em um problema (KORB et al., 2010). Um laboratório deseja estudar as relações entre as principais causas e efeitos de câncer no pulmão. Considerando que as duas causas principais que afetam as chances de um paciente desenvolver a doença, são: exposição a altos níveis de poluição e ser fumante. Uma vez que o paciente seja diagnosticado com câncer, o exame de Raio-*X*, geralmente, tem resultado positivo e o paciente apresenta certa dificuldade na respiração, chamada de dispneia. Ordenando as variáveis com base na opinião de um especialista, uma possível Rede Bayesiana foi construída, conforme Figura 13.

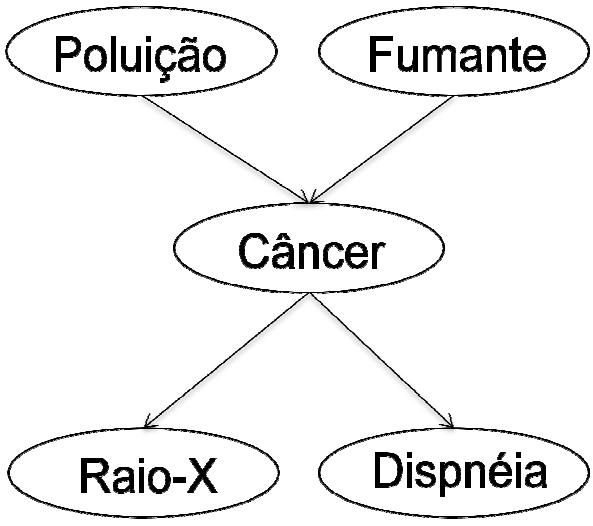

Figura 13 – Rede Bayesiana construída para o problema de câncer de pulmão
(Adaptado de: KORB et al., 2010)

Cada um dos cinco nós dessa rede pode assumir dois estados (Tabela 6). Dessa forma, com a estrutura construída, e usando como base um banco de dados com registros de vários pacientes, é possível calcular as tabelas de distribuição de probabilidade conjunta para as variáveis (Figura 14).

Tabela 6 – Variáveis e seus possíveis estados no problema do câncer de pulmão

Variável	Possíveis estados
Poluição (*P*)	Baixo nível (*B*) ou Alto nível (*A*)
Fumante (*F*)	Verdadeiro (*V*) ou Falso (*F*)
Câncer (*C*)	Verdadeiro (*V*) ou Falso (*F*)
Raio-X (*X*)	Verdadeiro (*V*) ou Falso (*F*)
Dispnéia (*D*)	Verdadeiro (*V*) ou Falso (*F*)

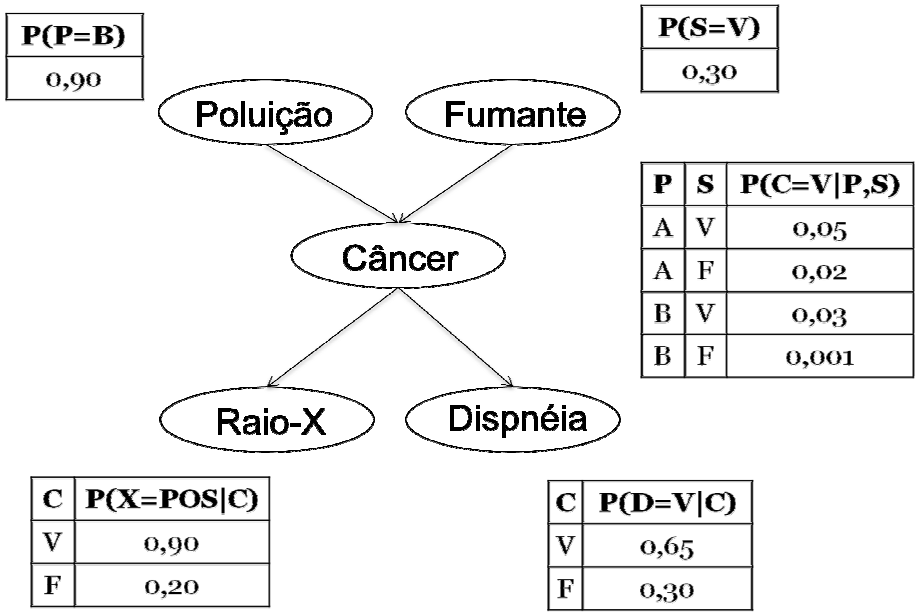

Figura 14 – Tabelas de distribuição de probabilidade conjunta para o problema de câncer no pulmão
(Adaptado de: KORB et al., 2010)

A fim de obter mais informações em uma Rede Bayesiana, é possível realizar a inferência diagnóstica, analisando a estrutura no sentido contrário aos arcos. Por exemplo, para calcular a probabilidade de um indivíduo ser fumante (*S=V*), dado que possui dispneia (*D=V*), é feito pela aplicação da Regra de Bayes, conforme equação a seguir:

$$P(S = V|D = V) = \frac{P(D = V|S = V)P(S = V)}{P(D = V)} \tag{23}$$

Cada termo da equação (23) deve ser calculado separadamente, com os valores obtidos pelas tabelas de DPC da Figura 14. Como a variável *Câncer* está entre os nós *Dispneia* e *Fumante*, o cálculo do termo P(*D=V*|*S=V*) precisa levar em consideração a probabilidade de um indivíduo ter câncer (*C=V*), ou não (*C=F*):

$$\begin{aligned} P(D = V|S = V) = P(D = V|C = V).P(C = V|S = V) + \\ + P(D = V|C = F).P(C = F|S = V) \end{aligned} \tag{24}$$

Porém, o nó *Câncer* também recebe influência do nó *Poluição*. Sendo assim, o cálculo do termo *P(C=V|S=V)* também considera a probabilidade do paciente ter sido exposto a altos níveis de poluição (*P=A*), ou não (*P=B*):

$$\begin{aligned} P(C = V|S = V) &= P(C = V|P = A, S = V).P(P = A) \\ &\quad + P(C = V|P = B, S = V).P(P = B) \\ P(C = V|S = V) &= 0{,}05.0{,}1 + 0{,}03.0{,}9 \\ P(C = V|S = V) &= 0{,}032 \end{aligned} \tag{25}$$

O termo *P(C=F|S=V)* pode ser encontrado usando o resultado da equação (25):

$$\begin{aligned} P(C = F|S = V) &= 1 - P(C = V|S = V) \\ P(C = F|S = V) &= 1 - 0{,}032 \\ P(C = F|S = V) &= 0{,}968 \end{aligned} \tag{26}$$

Com esses valores é possível encontrar o resultado da equação (24):

$P(D = V|S = V)$ = *0,65.0,032+0,30.0,968*
$P(D = V|S = V) = 0{,}311$

O próximo passo é calcular a probabilidade de ter dispneia. Esse cálculo também engloba os possíveis estados da variável *Câncer*, conforme a equação a seguir:

$$P(D = V) = P(D = V|C = V).P(C = V) + P(D = V|C = F).P(C = F) \quad (27)$$

Para o cálculo da probabilidade de ter câncer, é necessário considerar também as variáveis *Poluição* e *Fumante*:

$$\begin{aligned} P(C = V) = {} & P(C = V|P = A, S = V).P(P = A).P(S = V) + \\ & +P(C = V|P = A, S = F).P(P = A).P(S = F) + \\ & +P(C = V|P = B, S = V).P(P = B).P(S = V) + \\ & +P(C = V|P = B, S = F).P(P = B).P(S = F) \end{aligned} \quad (28)$$

$P(C = V) = 0{,}0116$

Com esses termos calculados, é possível obter o resultado da equação (27):

$P(D = V) = P(D = V|C = V).P(C = V) + P(D = V|C = F).P(C = F)$
$P(D = V) = 0{,}65.0{,}0116 + 0{,}30.(1 - 0{,}0116)$
$P(D = V) = 0{,}304$

Com todos os termos da equação (23), a probabilidade de um indivíduo ser fumante, dado que ele possui dispneia, é:

$$P(S = V|D = V) = \frac{0{,}311.0{,}30}{0{,}304} = 0{,}307$$

3.5.3. Aprendizagem e Construção de uma Rede

A aprendizagem Bayesiana visa fornecer uma estrutura que melhor represente o problema e que facilite a obtenção de informações. Esse processo pode ser dividido em duas partes. Na primeira parte ocorre a aprendizagem da topologia da rede, ordenando as variáveis do problema e suas relações de causalidade. Na segunda parte acontece a aprendizagem dos parâmetros numéricos, quando é feito o cálculo das probabilidades condicionais.

Um especialista pode analisar e definir as duas etapas da aprendizagem Bayesiana, tomando por base apenas seu conhecimento prévio. Porém, tanto a estrutura da rede, quanto as tabelas de distribuição de probabilidade, podem ser obtidas através de um conjunto de dados. Nesse caso, um algoritmo é usado para gerar a Rede Bayesiana de forma automática (GONÇALVES, 2017).

3.5.4. Exemplo de Aplicação

Uma Rede Bayesiana pode ser definida como a representação compacta da distribuição de probabilidade conjunta do domínio de um problema. Para um especialista, essas estruturas mostram de forma simples e gráfica as relações de causalidade das variáveis que compõe um sistema (MARQUES et al., 2003). O exemplo de aplicação de Redes Bayesianas a seguir, mostra uma das diversas possibilidades em representar o problema abordado.

Considerando o processo seletivo de uma companhia, onde é necessário contratar funcionários com alto nível de inteligência, optou-se por selecionar alunos recém-formados. Como não há uma forma direta de testar a inteligência dos candidatos, a companhia decidiu efetuar uma análise baseando-se em três critérios: a nota em uma disciplina específica, coeficiente de rendimento e uma carta de recomendação.

Pelos critérios escolhidos, algumas observações podem ser feitas. A nota de um aluno em uma disciplina específica, e de interesse para a empresa, depende da inteligência do estudante e da dificuldade em cursar a matéria. O coeficiente de rendimento no decorrer da faculdade também depende da inteligência do aluno. No caso da carta de recomendação, é provável que o professor não se lembre do desempenho de todos os seus alunos, logo, a carta pode ser redigida com base nas notas dos estudantes. No total, seis variáveis e seus possíveis estados podem ser listados, conforme Tabela 7. Com a seleção das variáveis relevantes para o

domínio do problema, é necessário ordená-las de acordo com suas relações de causalidade. A Figura 15 mostra uma possível estrutura de Rede Bayesiana para esse exemplo.

Tabela 7 – Variáveis e possíveis estados do problema *Contratar*

Variáveis	Possíveis Estados
Nível de inteligência do aluno (NI)	Alto ou Baixo
Nota em determinada disciplina (Nota)	Alta ou Baixa
Nível de dificuldade na disciplina (DIF)	Alto ou Baixo
Coeficiente de Rendimento (CR)	Alto ou Baixo
Carta de recomendação (Carta)	Forte ou Fraca
Contratar	Sim ou Não

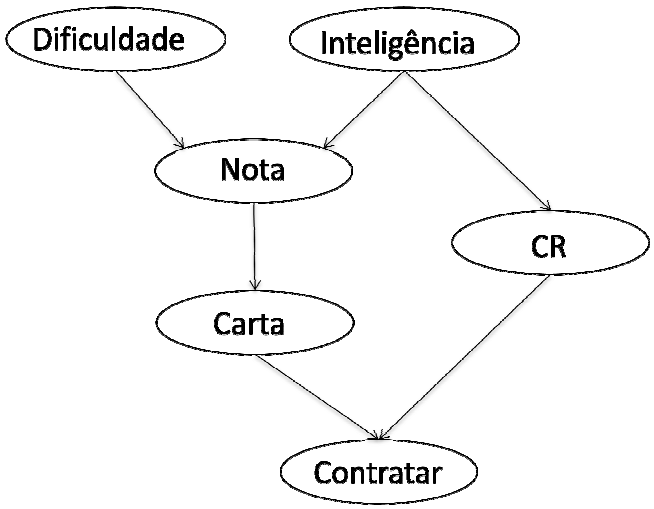

Figura 15 – Estrutura de Rede Bayesiana selecionada para o problema *Contratar*

Com a estrutura definida, é necessário montar as tabelas de distribuição de probabilidade conjunta. Nesse exemplo, os valores escolhidos simulam casos onde há um conjunto de dados para ser usado no cálculo dessas probabilidades. A Figura 16 mostra as seis tabelas para a estrutura da Rede Bayesiana gerada.

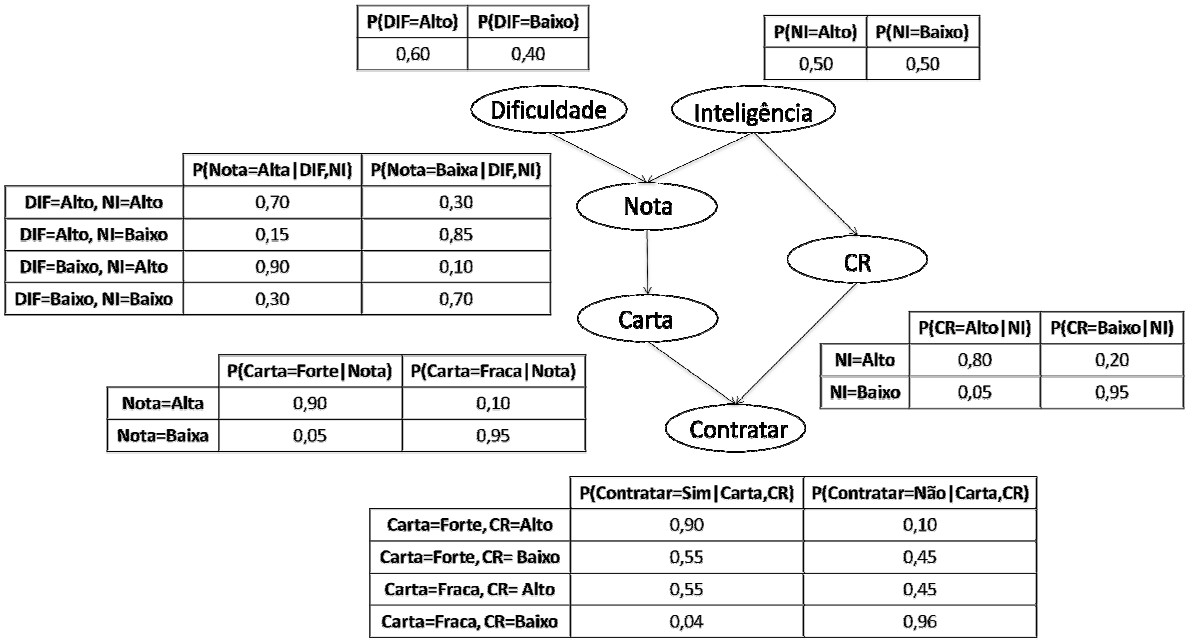

Figura 16 – Tabelas de distribuição de probabilidade conjunta do exemplo *Contratar*

Os nós raízes *Dificuldade* e *Inteligência* possuem tabelas com probabilidades à priori, pois não há nós que exerçam influência sobre eles. A variável *CR* é influenciada apenas pela variável *Inteligência*, já que um bom desempenho acadêmico (*CR*=Alto) está ligado a um bom nível de inteligência do aluno (*NI*=Alto). A variável *Carta* recebe influência apenas da variável *Nota*, mostrando que há alta probabilidade de um aluno ter uma carta com fortes recomendações, dado que foi observada uma nota alta em determinada disciplina.

As variáveis *Nota* e *Contratar* são influenciadas por dois nós pais, apresentando assim, tabelas com mais probabilidades a serem definidas. A variável *Nota* mostra que há probabilidade de 0,90 de um aluno ter bom desempenho em uma disciplina (*Nota*=Alta), dado que a matéria não é difícil (*DIF*=Baixo) e ele possui alto nível de inteligência (*NI*=Alto).

Já a variável *Contratar* recebe influência dos nós *Carta* e *CR*. Através dos diferentes estados que os nós pais podem assumir, é possível observar a alta probabilidade existente em duas situações. Um aluno com carta de recomendação forte e alto *CR*, possui probabilidade de 0,90 de ser contratado. Já um aluno com carta de recomendação fraca e baixo *CR*, possui probabilidade de 0,96 em não ser contratado.

Dessa forma, as Redes Bayesianas permitem inferir sobre o domínio de um problema, representando graficamente e quantificando as relações entre suas variáveis. Vale ressaltar que a estrutura feita para esse exemplo é apenas uma das formas de ordenar os nós.

3.5.5. Vantagens e Desvantagens das Redes Bayesianas

Sendo um método que utiliza o raciocínio probabilístico, as Redes Bayesianas permitem tomar decisões mesmo com grande quantidade de dados e informações insuficientes. Também permite expressar as relações causais entre as variáveis de forma visual e de fácil entendimento. Outra vantagem do método é a apresentação de tabelas de probabilidade condicional compactas, já que cada variável recebe influência apenas dos seus nós pais.

Como desvantagem, o algoritmo de Redes Bayesianas exige um bom conhecimento do problema para construir uma base de dados probabilística, mesmo com informações incompletas, o que pode exigir gastos. Outra desvantagem está na eliminação de dados para compactar o problema, o que pode eliminar parâmetros importantes.

3.6. Algoritmos Genéticos

Inspirados na teoria de seleção natural das espécies, proposta por Darwin, os algoritmos genéticos são técnicas usadas para busca e otimização de problemas complexos. A estratégia usada por esses algoritmos é a geração de uma população inicial composta por indivíduos que representam as possíveis soluções do problema. Esses indivíduos são codificados em estruturas chamadas de cromossomos que passam pelas gerações e evoluem de acordo com o princípio de seleção e sobrevivência dos mais aptos.

Na natureza, observa-se a competição de indivíduos por recursos básicos à sobrevivência. Os indivíduos que não tem sucesso em obter esses recursos possuem uma probabilidade menor em ter seus genes transferidos para as próximas gerações, e consequentemente, tem menos chance de deixar descendentes. Já os indivíduos que tem sucesso, possuem uma probabilidade maior em se manter nas próximas gerações, e dessa forma, produzir novos indivíduos com características mais adequadas ao seu meio ambiente. De forma análoga, a população de indivíduos representa o espaço de busca que contém possíveis soluções. As gerações são representadas pelos ciclos e o meio ambiente é o problema a ser resolvido (ROSA et al., 2009).

Todos os indivíduos da população são avaliados por uma função e recebem uma medida de aptidão, que reflete o quão boa uma solução é para o problema. Para que ocorra a geração de descendentes, um conjunto de indivíduos é selecionado com base na sua aptidão,

que serão, posteriormente, submetidos aos operadores de cruzamento (*crossover*) ou mutação. Esse processo é repetido até que o critério de parada seja atingido.

Dessa forma, os algoritmos genéticos otimizam problemas fornecendo a melhor solução possível de acordo com a aplicação desejada, mas não garante que haja convergência para uma solução ótima. A Figura 17 mostra um fluxograma com as principais etapas desse método.

Como vantagens, os Algoritmos Genéticos são robustos, podem ser usados em conjunto com outras técnicas e ser aplicados em diversos tipos de problemas. Como desvantagens, esse método apresenta dificuldade em encontrar a ótima solução exata e também possui a necessidade em ter um grande número de avaliações de função de aptidão, ocasionando em um desempenho mais lento (PINHO et al., 2013; LACERDA et al., 1999).

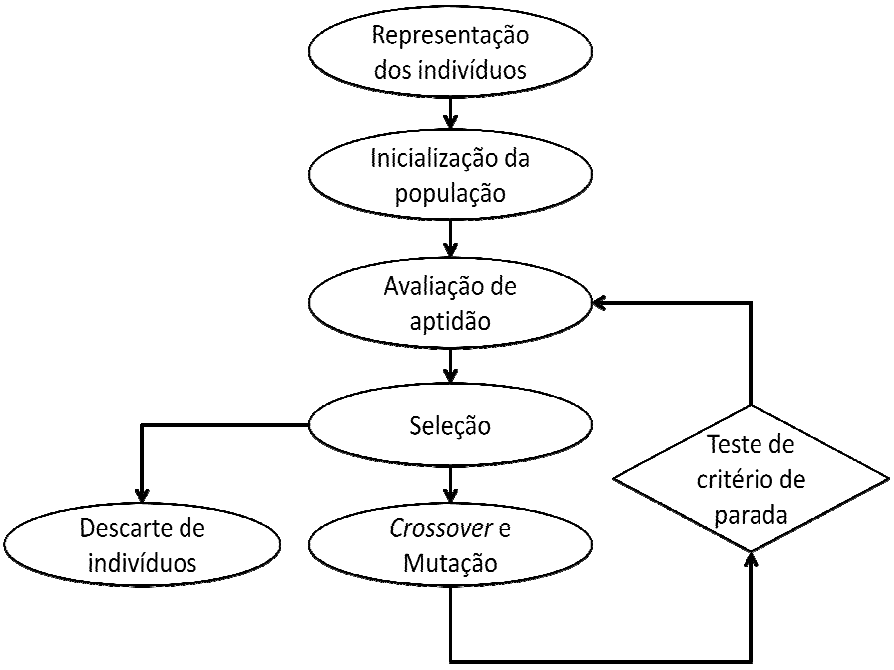

Figura 17 – Fluxograma básico de um algoritmo genético
(Adaptado de: ROSA et al., 2009)

4. MODELO PROPOSTO

O modelo desenvolvido neste projeto tem como objetivo aprimorar a detecção de anormalidades respiratórias, decorrentes da fibrose cística, por meio das características fornecidas pela FOT. Outro objetivo é a geração de estruturas de Redes Bayesianas que melhor descrevam as relações existentes entre essas características.

Inicialmente, os dados da FOT podem passar pelo processo de seleção de atributos. Em seguida, essas características selecionadas também podem passar pelo método do produto cruzado, quando é gerado um novo conjunto de dados composto por colunas extras com o cálculo do produto dessas características. O exemplo descrito na equação (29) mostra a saída resultante da aplicação do produto cruzado em um conjunto de dados (X), inicialmente composto por dois atributos, X_1 e X_2. Dessa forma, é possível apresentar ao modelo dados em uma dimensão mais alta, como tentativa de aumentar sua acurácia[1].

$$X = X_1^2 + X_1X_2 + X_2^2 \tag{29}$$

Com os atributos de entrada definidos, o conjunto de dados é submetido a quatro modelos de algoritmos de aprendizado de máquina: 1-*Nearest Neighbor* (1-NN), *Adaboost* (ADAB), *Random Forest* (RF) e *Radial Support Vector Machine* (RSVM). Esses dados também são submetidos às Redes Bayesianas sintetizadas por algoritmo genético (RBGAOT), usado para construção e seleção de estruturas, capazes de fornecer maior interpretabilidade das características mecânicas do sistema respiratório. Antes de passar pelo RBGAOT, os dados precisam ser discretizados a fim de viabilizar o cálculo das distribuições de probabilidade usado nesse método.

No presente estudo, todos os algoritmos de aprendizado de máquinas fornecem uma classificação como resultado final e tem seu desempenho calculado por meio da AUC. No caso do RBGAOT, também são obtidas estruturas que apresentaram melhor resultado durante a classificação e podem ser analisadas. Todo o modelo proposto, desenvolvido no *software* Matlab R2016a, foi descrito nos itens a seguir, de acordo com seu fluxograma resumido (Figura 18).

---

[1] Disponível em: <https://jakevdp.github.io/PythonDataScienceHandbook/05.04-feature-engineering.html>, Acessado em: 19/06/2018.

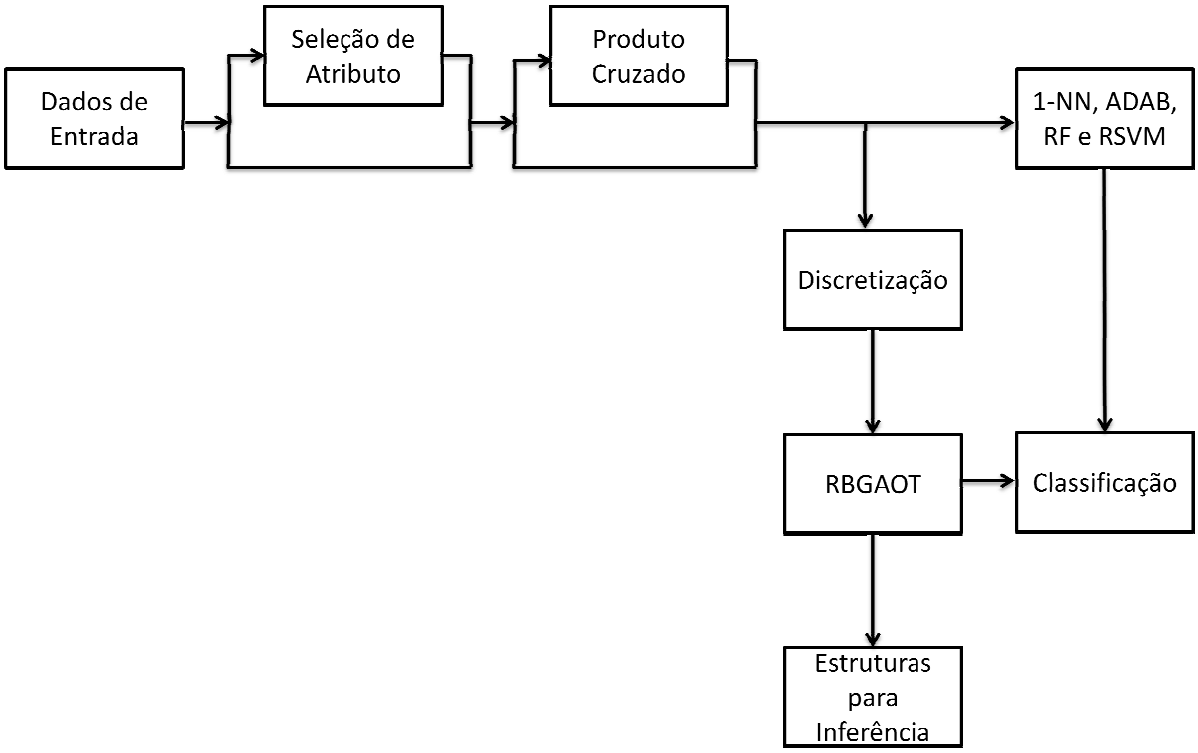

Figura 18 – Fluxograma resumido do modelo proposto

4.1. Dados de Entrada

O conjunto de dados usado neste projeto foi obtido através de exames realizados por um sistema de técnica de oscilações forçadas (FOT), desenvolvido no Laboratório de Instrumentação Biomédica da UERJ (LIB-UERJ). O procedimento para realizar os exames pela FOT em cada indivíduo, consistiu em três medições com intervalo de um minuto e duração de 16 segundos. A fim de evitar o vazamento do ar e induzir a respiração normal pelo bocal do equipamento, foi necessário que os indivíduos fizessem uso de *clip* nasal e permanecessem sentados durante o exame (MIRANDA et al., 2013).

O aparelho usado forneceu a impedância do sistema respiratório em uma faixa de frequências de 4 a 32Hz, que foi medida com incrementos de 2Hz. Através de um bocal, um alto-falante gerou oscilações de pressão com amplitude de 1 cmH_2O no sistema respiratório do paciente, durante sua respiração espontânea. Um pneumotacômetro e um transdutor de pressão foram usados para medir esses sinais de fluxo e de pressão, respectivamente, próximos à boca do paciente (LIMA et al., 2015).

4.2. Seleção de Atributos

Durante o projeto de construção de um classificador, a seleção de atributos de entrada é aplicada com o intuito de escolher as características que melhor descrevem o problema e, dessa forma, melhorar o desempenho do algoritmo. O uso dessa estratégia também permite a redução da complexidade do modelo, já que ocorre uma redução nos atributos e, consequentemente, uma diminuição do número de parâmetros que precisa ser estimado. Outras vantagens da técnica são o aumento da velocidade de execução do algoritmo, a possibilidade em visualizar os dados e a obtenção de uma melhor compreensão do processo que gera os resultados obtidos (GUYON et al., 2003).

Há duas formas principais de realizar a seleção de atributos. A primeira forma é através de um especialista que destaca os parâmetros que melhor descrevem o problema, com base em sua própria experiência. O outro método é feito de forma automática e pode utilizar técnicas de filtragem, *Wrapper* ou o método embutido. A filtragem realiza a classificação ordenada das características antes que os dados sejam submetidos ao algoritmo. Essa ordenação é feita com base em um critério escolhido, como coeficientes de correlação ou testes estatísticos. A técnica de *Wrapper* também é usada antes da classificação, fazendo uso de algoritmos de aprendizado de máquina para avaliar subconjuntos criados a partir do conjunto de dados (HORTA et al., 2010). O subconjunto que apresentar melhor desempenho tem seus atributos selecionados para o classificador. Já o método embutido realiza a seleção de variáveis durante o treinamento. Ele é usado especificamente em alguns algoritmos, como as árvores de decisões, que selecionam atributos no seu próprio processo de construção de um modelo (AMARAL et al., 2013). O RBGAOT também pode ser considerado um caso de algoritmo com seleção embutida, visto que mesmo submetendo um conjunto de atributos às Redes Bayesianas, variáveis podem ser descartadas durante a construção das estruturas geradas.

Neste projeto, o método *Wrapper* foi usado para a seleção de atributos por ser um método heurístico e guiado em sua busca pelo conjunto de atributos que maximize a média da AUC. Essa busca foi realizada de forma direta (*forward*), onde os atributos são acrescentados um por vez com base em um critério, até completar o subconjunto. O critério escolhido foi a taxa de acerto no algoritmo *K*-NN, com *K* igual a 1 (1-NN). O treinamento do classificador 1-NN foi feito através da validação cruzada *leave-one-out*, onde uma amostra *n* é testada com base nas *n*-1 amostras restantes (RODRIGUES et al., 2017). A seleção feita pelo 1-NN foi

implementada pela função *featself* disponível na *toolbox Pattern Recognition* (prtools) do *software* Matlab (DUIN, 2007).

Como o uso da AUC é recomendado em diagnósticos médicos (METZ, 1978; HANLEY et al., 1982), essa medida também foi testada como critério de seleção de atributos. Entretanto, as variáveis selecionadas por esse método não apresentaram desempenho superior ao das variáveis selecionadas pelo algoritmo 1-NN durante os experimentos. A seleção de atributos também foi feita pela consulta a um especialista, que elegeu o mesmo conjunto de variáveis obtido pelo algoritmo 1-NN.

4.3. Treinamento do Modelo

A técnica de validação cruzada foi usada durante o treinamento do modelo. Devido a pouca quantidade de amostras disponíveis e a grande quantidade de atributos fornecidos pela FOT, optou-se pelo método de validação cruzada por *k*-pastas, onde uma parte dos dados é separada para treino e o restante é destinado para teste do modelo. Durante esse processo, os dados são divididos em *k* pastas, sendo geralmente uma pasta usada para testar e *k*-1 pastas usadas para o aprendizado do modelo.

Esse processo é repetido *k* vezes e a cada iteração são usadas diferentes pastas para o conjunto de treino e teste, gerando então, *k* medidas de erro. A média desses *k* erros é a medida de generalização do modelo. Dessa forma, é possível evitar uma estimativa muito otimista fornecida por alguma das *k* partições, como poderia ocorrer na validação cruzada feita pela técnica *hold-out* (HASTIE, 2008). No exemplo da Figura 19 é possível observar a divisão para *k* igual a 10, valor escolhido para o uso da validação cruzada neste projeto.

A métrica usada para a seleção dos classificadores e a configuração escolhida para cada modelo estão descritas dos itens 4.4 a 4.6.

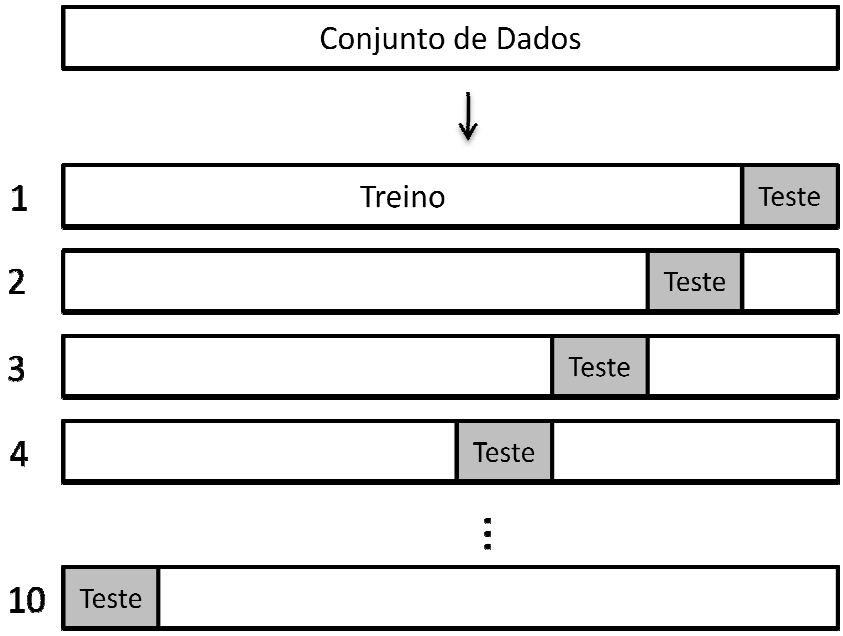

Figura 19 – Divisão para validação cruzada com 10 pastas

4.4. Medida de desempenho

A medida de desempenho usada para a seleção dos melhores modelos de classificação foi feita com base na área sob a curva ROC. A AUC é uma ferramenta normalmente usada para diagnóstico médico (METZ, 1978; HANLEY et al., 1982), que também fornece informações sobre a eficácia de algoritmos de aprendizado de máquina (HUANG et al, 2005).

Considerando um conjunto de dados com duas classes, positiva (*p*) e negativa (*n*), os rótulos obtidos na classificação desses dados podem ser respectivamente, *P* e *N*. A matriz confusão da Figura 20 fornece quatro possibilidades para uma instância ao ser classificada. Se a instância for positiva e classificada como positiva, é um caso de verdadeiro positivo (*VP*). Se for classificada como negativa pelo algoritmo, é um caso de falso negativo (*FN*). Se a instância for negativa e classificada como negativa, é um caso verdadeiro negativo (*VN*). Se for classificada como positiva, é um caso de falso positivo (*FP*) (FAWCETT, 2006).

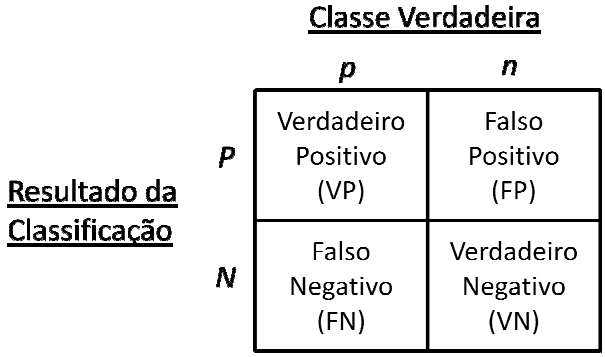

Figura 20 – Matriz confusão das possíveis classificações de uma instância
(Adaptado de: FAWCETT, 2006)

Pelos valores da matriz na Figura 20, podem ser calculadas diversas métricas, sendo a diagonal principal as classificações corretas do modelo. Para este projeto, além da AUC, também foram usadas a sensibilidade e a especificidade. A equação (30) mostra o cálculo da sensibilidade, que corresponde à probabilidade de ter uma classificação positiva quando a instância é positiva. Já a equação (31), mostra a probabilidade de ter uma classificação negativa quando a instância é negativa.

$$\text{Sensibilidade} = \frac{\text{VP}}{\text{VP} + \text{FN}} \tag{30}$$

$$\text{Especificidade} = \frac{\text{VN}}{\text{VN} + \text{FP}} \tag{31}$$

Para construir a curva ROC, são usados os valores de sensibilidade no eixo *y* e o resultado da subtração de 1-especificidade no eixo *x*, caracterizando a relação entre os benefícios (verdadeiros positivos) e os custos (falsos positivos) de um modelo (FAWCETT, 2006). Caso não seja possível selecionar visualmente um classificador, a área sob a curva ROC é calculada e, pelo seu valor, é possível definir o classificador com melhor desempenho.

4.5. Classificadores

Dentre os classificadores descritos no capítulo 3, quatro foram escolhidos e implementados pela *toolbox* prtools[2], com base em trabalhos realizados nessa linha de pesquisa (AMARAL et al., 2013; AMARAL et al., 2015; AMARAL et al., 2017): *K-Nearest Neighbor* (K-NN), *Adaboost* (ADAB), *Random Forest* (RF) e *Radial Support Vector Machine* (RSVM), cujos parâmetros foram definidos de acordo com o desempenho durante a validação cruzada.

No algoritmo *K*-NN, o valor de *K* foi definido de acordo com o erro encontrado durante o treinamento usando a área sob a curva ROC (AUC) como medida de desempenho (E_{AUC}). Foram avaliados diferentes valores de *K*, sendo *K* igual a 1 (1-NN) a configuração que apresentou melhor desempenho, conforme Tabela 8.

Tabela 8 – Resultados do treinamento do algoritmo K-NN

K	E_{AUC}
1	0,1252
3	0,1768
5	0,1467
7	0,1554
9	0,1746
11	0,1807
13	0,1866
15	0,1582
17	0,1501
19	0,1563

A quantidade de árvores de decisões usadas como classificadores simples no algoritmo *ADAB*, foi selecionada de acordo com o erro de AUC encontrado durante o treinamento. Conforme Tabela 9, o número de árvores de decisões usado que apresentou melhor desempenho foi 200.

---

[2] Prtools: *Toolbox for Pattern Recognition*. Disponível em: <http://prtools.org/>, Acessado em: 16/03/2018.

Tabela 9 – Resultados do treinamento do algoritmo ADAB

Número de Árvores de Decisão	E_{AUC}
50	0,1238
100	0,1296
150	0,1192
200	0,1019
250	0,1120

O algoritmo RF foi implementado de acordo com o erro encontrado durante o treinamento, utilizando a AUC como medida do desempenho do modelo testado. Nesse caso, foram avaliadas a quantidade de subgrupos de atributos formados e a quantidade de árvores geradas. De acordo com a Tabela 10, a configuração que apresentou menor erro possui 50 árvores geradas e tamanho do subconjunto de atributos igual a 1.

Tabela 10 – Resultados do treinamento do algoritmo RF

Árvores geradas / Subgrupos	10	20	50	100	150
1	0,1664	0,1152	0,0970	0,1268	0,1361
2	0,1245	0,1383	0,1030	0,1375	0,1139
3	0,1190	0,1363	0,1155	0,1235	0,1079
5	0,1472	0,1304	0,1147	0,1142	0,1219
7	0,1401	0,1246	0,1257	0,1148	0,1053

No caso do algoritmo RSVM, foi necessário definir dois parâmetros: o desvio padrão da base radial (*r*) e o parâmetro de regularização (*C*). A busca por esses parâmetros foi realizada por uma validação cruzada interna durante o treinamento[3].

Com o intuito de extrair informações a respeito das relações entre os atributos para obter uma explicação da classificação, utilizou-se um classificador baseado em Redes Bayesianas. De acordo com trabalhos já realizados (TONDA et al., 2012; LARRAÑAGA et al., 1996), é possível fazer uso de algoritmos evolutivos para o aprendizado dessas redes.

---

[3] Disponível em: <http://www.37steps.com/prhtml/prtools/rbsvc.htm>, Acessado em: 16/03/2018

No artigo "*Bayesian Network Structure Learning from Limited Datasets through Graph Evolution*" (TONDA et al., 2012), é proposto o uso de um algoritmo evolutivo para a aprendizagem das estruturas de Redes Bayesianas, tomando por base um conjunto de dados com número de amostras limitado. Esse trabalho apresenta a possibilidade de trabalhar usando diretamente estruturas gráficas. Sua função *fitness* é baseada na métrica de informação *Akaike*, considerando a precisão e a complexidade da estrutura fornecida durante o treinamento do modelo.

Já o artigo (LARRAÑAGA et al., 1996), mostra o uso de algoritmo genético para realizar a busca da melhor estrutura de Rede Bayesiana, com base em um conjunto de dados. Esse trabalho usa matrizes para representar suas soluções e propõe o uso de um operador responsável por corrigir redes que não forem DAG. A função *fitness* calcula a métrica *K2*, onde a estrutura com maior valor de probabilidade conjunta, dado um conjunto de treinamento, é selecionada.

Com base nesses trabalhos, a técnica de algoritmos genéticos foi utilizada para realizar a busca pela estrutura com melhor desempenho. A descrição desse método escolhido está nos itens a seguir.

4.6. Redes Bayesianas sintetizadas com Algoritmos Genéticos

As Redes Bayesianas foram implementadas pela *toolbox Probabilistic Graphical Model* 9.2.3 (PGM[4]), onde a primeira etapa consiste na leitura de um grafo direcionado e acíclico (MENSXMACHINA, 2011). No *software* Matlab essas estruturas são representadas por matrizes esparsas binárias, onde elementos iguais a 0 são eliminados e os elementos iguais a 1 representam as ligações entre as variáveis do problema, conforme Figura 21:

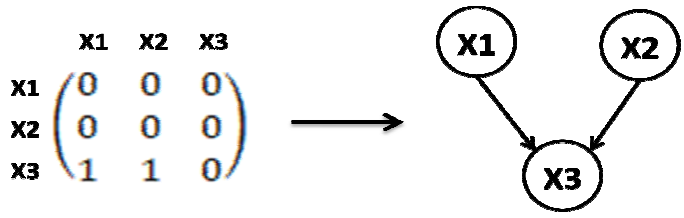

Figura 21 – Representação de uma Rede Bayesiana em matriz esparsa

[4] *Toolbox* PGM: *Probabilistic Graphical Model* 9.2.3. Disponível em: <http://mensxmachina.org/en/software/pgm-toolbox/>, Acessado em: 16/03/2018

Durante essa primeira etapa, podem ser apresentadas ao algoritmo estruturas inválidas de Redes Bayesianas, como redes que não sejam DAG ou redes que não possuam a variável classe. Essas matrizes são identificadas através do ajuste de linhas e colunas, realizado pela *toolbox* PGM, retornando redes com apenas uma linha. Dessa forma, elas são identificadas e recebem valores de AUC igual à zero, para que sejam descartadas durante as próximas iterações.

Em seguida, o aprendizado das distribuições de probabilidade conjunta (DPC) é feito pelo algoritmo BDeu (*Bayesian Dirichlet Equivalent Uniform*), desenvolvido por Heckerman (ONISKO et al., 2001) e baseado na métrica *Bayesian Dirichlet* desenvolvida por Cooper e Herskovits (COOPER et al., 1991). Essa pontuação corresponde ao logaritmo da probabilidade à posteriori de uma rede B_s, dado um conjunto de dados *A*, logo, a pontuação é obtida pelo cálculo de $\log(P(B_s|A))$. A métrica BDeu deve corresponder à capacidade de uma rede em capturar a probabilidade conjunta dos dados e prever novas amostras, apresentando então, sua relação direta com a capacidade de inferência da estrutura analisada (BROWN et al., 2004).

Durante essa segunda etapa, as tabelas de DPC são calculadas e contém: o nome da variável analisada, os diversos valores que seus respectivos nós pais podem assumir e as probabilidades condicionadas aos nós pais. No exemplo da Tabela 11, a variável analisada é *X3*, os valores que as variáveis podem assumir são positivo (*p*) ou negativo (*n*) e os nós pais sãos *X1* e *X2*. Com a estrutura definida e as tabelas de DPC calculadas, a função *bayesnet* realiza a construção da Rede Bayesiana.

Tabela 11 – Exemplo de tabela de DPC

	P(X3=p|X1,X2)	*P(X3=n|X1,X2)*
X1=p, X2=p	0,75	0,25
X1=n, X2=p	0,05	0,95
X1=p, X2=n	0,85	0,15
X1=n, X2=n	0,10	0,90

A terceira etapa disponível na *toolbox* PGM é um mecanismo de inferência que pode ser usado para a classificação de um conjunto de teste, através de uma Rede Bayesiana já construída. O algoritmo *Junction Tree* (BARBER, 2003) é aplicado com o intuito de fornecer uma ótima sequência de decomposição, ou marginalização, dessa rede. A estrutura resultante desse algoritmo é capaz de calcular a distribuição de probabilidade marginal de cada amostra

de teste. Essa distribuição corresponde às probabilidades de uma nova amostra pertencer a cada uma das classes do problema.

Considerando como exemplo um conjunto de dados com duas classes: positiva e negativa, durante a terceira etapa são calculadas as probabilidades marginais de uma amostra de teste pertencer à classe positiva ou a classe negativa. Esses valores são importantes para a classificação com Redes Bayesianas, sendo que a maior probabilidade marginal encontrada define o rótulo dessa amostra.

As três etapas descritas, são a base para o uso das Redes Bayesianas pela *toolbox* PGM. Porém, estratégias externas podem ser aplicadas para trabalhar com o conjunto de dados, gerar diversas estruturas e selecionar aquela que melhor descreve o problema. Neste trabalho, duas estratégias foram escolhidas: a discretização dos dados de entrada e a aplicação de um algoritmo genético no treinamento das redes.

4.6.1. Discretização dos Dados

As tabelas de distribuição de probabilidade conjunta de uma Rede Bayesiana quantificam as relações existentes entre suas variáveis, abordando os diversos estados que seus nós pais podem assumir. Quando o conjunto de treinamento é composto por valores contínuos, é necessário escolher uma abordagem para esses dados. Neste trabalho, foi escolhido o método de discretização para possibilitar o uso das Redes Bayesianas, devido sua simples implementação e interpretação na leitura das tabelas de distribuição de probabilidade conjunta, por parte da equipe médica.

A discretização pode ser definida como o processo de transformação de uma variável contínua em uma variável discreta. Dessa forma, as amostras passam a ser apresentadas em intervalos cujos limiares são pontos de corte que podem ser definidos por diversos tipos de cálculos. O uso de dados discretizados facilita até mesmo a análise de características do problema por parte dos usuários e especialistas.

Todos os métodos usados para discretização podem ser classificados como supervisionados e não supervisionados. No método supervisionado, as classes dos atributos são levadas em consideração, já no caso não-supervisionado, a discretização é feita considerando apenas os valores dos atributos (CARVALHO, 2010).

O método escolhido para discretizar os dados fornecidos pela FOT é supervisionado e usa o conceito de entropia. Esse método tem como objetivo determinar um ponto de corte que seja capaz de gerar os mais puros subconjuntos possíveis, com base no maior ganho de informação (FAYYAD et al., 1993):

$$Ganho(A) = E(conjunto\ Atual) - \sum E(subconjuntos) \tag{32}$$

Sendo:
E( ) a entropia
Conjunto atual os dados a serem discretizados
Subconjuntos os intervalos criados para o atributo *A*

De acordo com a equação (32), para aumentar o ganho de informação é necessário minimizar o somatório das entropias relativas aos subconjuntos (MERSCHMANN, 2007):

$$E(subconjunto) = -\sum_{i=1}^{C} p_i \log_2(p_i) \tag{33}$$

Sendo:
C a quantidade de classes do problema
p_i a probabilidade da classe *i* ocorrer

Os pontos de corte são calculados até que o critério de parada seja alcançado e haja uma diminuição dos intervalos em cada atributo. O princípio usado como critério de parada é o MDL (*Minimum Description Length*) que compara se a criação de um novo intervalo aumenta o ganho de informação. Quando não houver mais aumento, o ponto de corte que fornecer o maior ganho de informação é selecionado (FAYYAD et al., 1993).

4.6.2. RBGAOT

A estratégia escolhida para realizar a aprendizagem da estrutura das Redes Bayesianas foi o uso de algoritmos genéticos, por meio da *toolbox Genetic Algorithms for optimization* (GAOT). Para implementar o uso dessa técnica na criação e seleção da melhor estrutura que descreva as relações entre as variáveis do problema, foi feita a junção das *toolboxes* de Redes Bayesianas e Algoritmo Genético, chamada RBGAOT.

O RBGAOT gera de forma aleatória diversas redes representadas em matrizes de adjacência que apresentam possíveis soluções ao problema. Cada rede é construída com base nessas matrizes e tem suas distribuições de probabilidade calculadas pela *toolbox* de Redes Bayesianas (PGM). Uma vez que a rede já esteja com todas as suas características definidas, é possível analisar seu desempenho classificatório.

Há cinco elementos principais que precisam ser definidos para o uso do RBGAOT: representação do cromossomo, criação de uma população inicial, função de aptidão, função de seleção e operadores genéticos.

4.6.3. Representação do cromossomo

Um cromossomo equivale a cada indivíduo da população em um algoritmo genético e é composto por uma sequência de genes. No RBGAOT, um cromossomo corresponde à estrutura de uma Rede Bayesiana com *n* variáveis e genes formados por dígitos binários. Uma rede pode ser representada por uma matriz de adjacência *C* de tamanho *n* x *n*, cujos elementos são descritos de acordo com as ligações existentes, ou não, entre *j* e *i*, conforme a seguir:

$$c_{ij} = \begin{cases} 1, & se\ j\ é\ pai\ de\ i \\ 0, & se\ não\ existe\ ligação\ entre\ j\ e\ i \end{cases} \tag{34}$$

Dessa forma, as ligações entre as variáveis são expressas em uma matriz que, por sua vez, pode ser decomposta coluna a coluna para gerar um vetor (LARRAÑAGA et al., 1996), conforme o exemplo a seguir:

$$C = \begin{pmatrix} c_{11} & \cdots & c_{1n} \\ \vdots & \ddots & \vdots \\ c_{n1} & \cdots & c_{nn} \end{pmatrix}$$

$$Vetor = c_{11}c_{21}c_{31} \ldots c_{n1}\ c_{12}c_{22}c_{32} \ldots c_{n2} \ldots c_{1n}c_{2n}c_{3n} \ldots c_{nn}$$

4.6.4. População Inicial

Um indivíduo na população inicial corresponde a uma estrutura de Rede Bayesiana que pode ser selecionada para formar a próxima geração ou passar pelos operadores de mutação e *crossover*. A criação de uma população inicial P_{ij}, é feita de forma aleatória com uma distribuição uniforme (*U*), conforme a equação (35) (CIVICIOGLU, 2013):

$$P_{ij} = U(A_j, B_j), para \text{ i} = 1,2,\dots,N \text{ e j} = 1,2,\dots,D \tag{35}$$

Sendo:
A_j e B_j o limite mínimo e máximo de cada indivíduo do vetor *P*, respectivamente.
N o tamanho da população
D a dimensão dos indivíduos da população

No RBGAOT, a população inicial foi criada com 15 indivíduos formados por valores entre 0 e 1, e 20 gerações. Em seguida, toda a população teve seus genes aproximados para valores binários, conforme o sistema da equação (36):

$$P_{ij} = \begin{cases} 1, U(A_j, B_j) > 0{,}5 \\ 0, U(A_j, B_j) < 0{,}5 \end{cases} \tag{36}$$

4.6.5. Função de Avaliação

A função de avaliação ou *fitness* é usada para determinar a aptidão de cada indivíduo gerado durante a busca pela melhor solução, no RBGAOT. Cada vetor, que representa um indivíduo gerado, é recebido por essa função e convertido para uma matriz esparsa, conforme o processo descrito no item 4.6.3. Uma vez que se tenha a estrutura em formato de matriz, o RBGAOT faz uso da *toolbox* de Redes Bayesianas para o treinamento e teste da estrutura gerada.

As duas saídas fornecidas pela função de avaliação desse algoritmo são: o valor da AUC da estrutura testada e um vetor com a probabilidade das amostras do grupo de teste obtidas durante a classificação. Essas probabilidades serão usadas na construção da curva ROC.

4.6.6. Função de Seleção

O RBGAOT realiza a seleção de indivíduos de forma probabilística pelo método de roleta, onde os mais aptos têm maior probabilidade de serem escolhidos para formar a próxima geração. Também foi usado o *ranking* por normalização geométrica, para ordenação dos indivíduos (*i*) de acordo com a probabilidade *P(i)*, definida conforme equação (37). Essa técnica evita que indivíduos com aptidão muito acima da média sejam sempre escolhidos, levando o algoritmo a uma convergência prematura (HOUCK et al., 1995).

$$P(i) = q_t(1-q)^{r-1},\ q_t = \frac{q}{1-(1-q)^T} \tag{37}$$

Sendo:
i o indivíduo ou possível solução
q a probabilidade de selecionar o melhor indivíduo
r o *rank* do indivíduo (onde 1 é o melhor)
T o tamanho da população

4.6.7. Operadores Genéticos

Os operadores genéticos são mecanismos básicos de busca usados pelo algoritmo genético e tem como função criar novos indivíduos com base na população já existente. Um dos principais operadores é o *crossover*, que usa dois indivíduos pais para gerar dois novos indivíduos filhos através do cruzamento de seus cromossomos.

Apesar da *toolbox* GAOT disponibilizar diversos tipos de *crossover*, o cruzamento simples foi o que apresentou melhor desempenho, onde dois indivíduos pais (*X* e *Y*) de uma população de tamanho *m* formam dois novos indivíduos (*X'* e *Y'*). Para isso, um número aleatório *r* é criado por uma distribuição uniforme de 1 até *m*, para ser aplicado como ponto de corte, conforme as equações (38) e (39) (HOUCK et al., 1995).

$$x'_i = \begin{cases} x_i, & se\ i < r \\ y_i, & caso\ contrário \end{cases} \tag{38}$$

$$y'_i = \begin{cases} y_i, & se\ i < r \\ x_i, & caso\ contrário \end{cases} \tag{39}$$

Quanto maior for o valor da taxa de *crossover* escolhida, maior será a quantidade de novas estruturas promissoras, uma vez que ele combina as características de pais com alta aptidão. Entretanto, isso pode levar a uma convergência prematura da evolução. Caso a taxa de crossover seja muito baixa, o algoritmo poderá demorar a convergir para uma solução aceitável. Sendo assim, a taxa de 0,6 foi escolhida para o *crossover* no RBGAOT.

O operador de mutação também é muito usado nos algoritmos genéticos. Seu objetivo é alterar o cromossomo de um indivíduo *X* da população e gerar apenas uma nova solução *X'*. Dentre os operadores de mutação disponíveis na *toolbox* usada, a mutação binária apresentou melhor desempenho, realizando alterações com base em uma probabilidade (p_m). Os genes dos novos indivíduos são definidos conforme a equação (40) (HOUCK et al., 1995):

$$x'_i = \begin{cases} 1 - x_i, & se\ U(0,1) < p_m \\ x_i, & caso\ contrário \end{cases} \tag{40}$$

A taxa de mutação pode evitar que o algoritmo fique estagnado em uma solução, permitindo que a busca seja realizada em mais pontos no espaço de soluções. Valores mais altos tendem a tornar essa busca aleatória. Dentre os valores testados na faixa de 0,005 e 0,1, a taxa de mutação de 0,01 gerou os melhores resultados e foi o valor escolhido para o RBGAOT.