sid.inpe.br/mtc-m19/2013/06.12.18.15-TDI

MÉTODOS DE SISTEMAS DINÂMICOS E MINERAÇÃO DE DADOS PARA INTERPRETAÇÃO DE SINAIS NÃO LINEARES

Laurita dos Santos

Tese de Doutorado do Curso de Pós-Graduação em Computação Aplicada, orientada pelos Drs. Elbert Einstein Nehrer Macau, e Joaquim José Barroso de Castro aprovada em 08 de julho de 2013.

URL do documento original:
<http://urlib.net/8JMKD3MGP7W/3E9TAEB>

INPE
São José dos Campos
2013

PUBLICADO POR:

Instituto Nacional de Pesquisas Espaciais - INPE
Gabinete do Diretor (GB)
Serviço de Informação e Documentação (SID)
Caixa Postal 515 - CEP 12.245-970
São José dos Campos - SP - Brasil
Tel.:(012) 3208-6923/6921
Fax: (012) 3208-6919
E-mail: pubtc@sid.inpe.br

sid.inpe.br/mtc-m19/2013/06.12.18.15-TDI

MÉTODOS DE SISTEMAS DINÂMICOS E MINERAÇÃO DE DADOS PARA INTERPRETAÇÃO DE SINAIS NÃO LINEARES

Laurita dos Santos

Tese de Doutorado do Curso de Pós-Graduação em Computação Aplicada, orientada pelos Drs. Elbert Einstein Nehrer Macau, e Joaquim José Barroso de Castro aprovada em 08 de julho de 2013.

URL do documento original:
<http://urlib.net/8JMKD3MGP7W/3E9TAEB>

INPE
São José dos Campos
2013

Dados Internacionais de Catalogação na Publicação (CIP)

Santos, Laurita dos.

Sa59m Métodos de Sistemas Dinâmicos e Mineração de Dados para Interpretação de Sinais Não Lineares / Laurita dos Santos. – São José dos Campos : INPE, 2013.

xxvi + 121 p. ; (sid.inpe.br/mtc-m19/2013/06.12.18.15-TDI)

Tese (Doutorado em Computação Aplicada) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2013.

Orientadores : Dr. Elbert Einstein Nehrer Macau, e Joaquim José Barroso de Castro.

1. sinais não lineares. 2. sistemas dinâmicos. 3. mineração de dados. 4. séries temporais de vento neutro 5. séries temporais de intervalos RR I.Título.

CDU 519.246.8:004.8

Aprovado (a) pela Banca Examinadora em cumprimento ao requisito exigido para obtenção do Título de *Doutor(a)* em

Computação Aplicada

Dr. Lamartine Nogueira Frutuoso Guimarães

Presidente / INPE / SJCampos - SP

Dr. Elbert Einstein Nehrer Macaul

Orientador(a) / INPE / São José dos Campos - SP

Dr. Joaquim José Barroso de Castro

Orientador(a) / INPE / SJCampos - SP

Dr. Marcos Gonçalves Quiles

Membro da Banca / INPE / São José dos Campos - SP

Dr. Moacir Fernandes de Godoy

Convidado(a) / FAMERP / São José do Rio Preto - SP

Dr. Luís Antônio Aguirre

Convidado(a) / UFMG / Belo Horizonte - MG

Este trabalho foi aprovado por:

( ) maioria simples

(✓) unanimidade

Aluno (a): *Laurita dos Santos*

...“Como você pode ver, não são as respostas que movem o mundo, são as perguntas”.

Campanha Publicitária do Canal Futura, 2009

À minha ***Família****...*

AGRADECIMENTOS

Muito especialmente, desejo agradecer aos meus orientadores Prof. Dr. Elbert Einstein Nehrer Macau e Prof. Dr. Joaquim José Barroso de Castro, pela disponibilidade, atenção dispensada, paciência, dedicação e profissionalismo ... um Muito Obrigada.

Aos Colaboradores Prof. Dr. Moacir Fernandes de Godoy, Prof. Dr. Christophe Letellier e Prof. Dr. Ubiratan Santos Freitas, pela disponibilidade, atenção dispensada e apoio fundamental no desenvolvimento desta tese.

À CAPES pelo suporte financeiro e concessão da bolsa de Doutorado Sanduíche PDSE, processo número $8954 - 11 - 9$.

À minha família, em particular, aos meus pais por todo o apoio em todos os momentos, principalmente os mais difíceis.

Aos meus amigos e colegas do INPE pelo apoio, amizade e pela oportunidade de convivência.

Aos membros da banca examidora Prof. Dr. Lamartine Nogueira Frutuoso Guimarães, Prof. Dr. Marcos Gonçalves Quiles, Prof. Dr. Moacir Fernandes de Godoy e Prof. Dr. Luis Antônio Aguirre pelas valiosas contribuições.

RESUMO

Neste trabalho analisamos grupos de séries temporais de Variabilidade da Frequência Cardíaca de indivíduos de diversas condições clínicas, mas que podem apresentar semelhanças em termos de variabilidade; por exemplo, grupos de séries temporais de recém-nascidos prematuros e de recém-nascidos normais. Temos por objetivo principal discriminar o comportamento de sistemas semelhantes através dos grupos de séries temporais usando métodos de sistemas dinâmicos e mineração de dados. Os métodos de sistemas dinâmicos fornecem índices dos grupos de séries temporais e que são usados como padrões de entrada para as técnicas de mineração de dados. As técnicas usadas são os classificadores de árvore de decisão e máquinas de vetores de suporte. Para tal, usamos um conjunto de 514 séries temporais de intervalos RR oriundos de três diferentes bancos de dados. Uma pré-classificação do conjunto de séries temporais é realizada clinicamente por um médico especialista. Esse conjunto de séries temporais é pré-processado usando-se a filtragem convencional (do especialista) e a filtragem adaptativa que permite o pré-processamento de um grande volume de dados. Aplicamos o mesmo conjunto de métodos em outro estudo de caso, envolvendo também sinais não lineares. Esse segundo estudo de caso é relacionado à compreensão do comportamento da alta atmosfera e para tal, usamos séries temporais de vento neutro, oriundas do banco de dados do INPE. Esse conjunto é constituído de 47 séries temporais e que são classificadas em dois grupos. As principais contribuições deste trabalho estão relacionadas à filtragem adaptativa que é demonstrado ser estatisticamente equivalente ao processo de filtragem tradicional. Mostramos também que os métodos são capazes de detectar diferenças entre os sistemas. Além disso, verificamos que há um grupo específico de índices (parâmetros do mapa de primeiro retorno e medida da tendência central) que detectam diferenças entre os sistemas, conforme a comparação dos vários grupos de séries temporais.

DYNAMICAL SYSTEMS AND DATA MINING METHODS FOR INTERPRETING NONLINEAR SIGNALS

ABSTRACT

One way to study the dynamics of systems is to analyze time series obtained from these systems. For example, to study the variation of the nervous system of an individual, we can analyze the time series related to the heart rate variability. However, we are interested in the differences between systems that may exhibit similar dynamics. In this study we aim to characterize the systems dynamics with dynamics similar using time series methods for dynamic systems and data mining. To solve this task, we propose a methodology that links two areas of knowledge: dynamics systems and data mining methods. In this methodology, the methods of dynamics systems provide features of groups of time series. These measures are used as input patterns to the data mining techniques (decision tree classifiers and support vector machines). We applied this methodology in 514 RR intervals time serie. This methodology is employed in another case study, involving another set of non-linear signals. This second case study is related to understanding the dynamics of the upper atmosphere and for that, we can use time series of neutral wind (from INPE database). Neutral wind time series are detected in the upper atmosphere and are related to the detection of meteors entering the atmosphere. This case study consists of 47 series and are classified into two groups according to location of the wind measuring data. In both applications noted that the methods are capable of detecting differences between the systems. Moreover, we verified that a specific group of measures is able to detect differences between the systems, according to the comparison of the various classes of time series.

LISTA DE FIGURAS

Pág.

LISTA DE TABELAS

Pág.

LISTA DE ABREVIATURAS E SIGLAS

VFC	–	Variabilidade da Frequência Cardíaca
AR	–	Modelo Autoregressivo
EQM	–	Erro Quadrático Médio
CTM	–	Medida da Tendência Central
RQA	–	Análise de Quantificação de Recorrência
MD	–	Mineração de Dados
DS	–	Dinâmica Simbólica
LMC	–	Medida de Complexidade
SVM	–	Máquinas de Vetores de Suporte
SNA	–	Sistema Nervoso Autônomo
ECG	–	Eletrocardiograma
SNS	–	Sistema Nervoso Simpático
SNP	–	Sistema Nervoso Parassimpático
QRS	–	Complexo das Ondas QRS encontradas nos Eletrocardiogramas
NUTECC	–	Núcleo Transdisciplinar de Estudos de Complexidade e Caos
CORIA	–	*Complexe de Recherche Interprofessionel en Aerothermochimie*
PhysioNet	–	Banco de Dados para Sinais Fisiológicos
RNP	–	Recém-Nascido Prematuro
RNN	–	Recém-Nascido Normal
VOL	–	Adulto Jovem Saudável
ADC	–	Adulto sob Dieta de muito baixo valor calórico
PC	–	Adulto Coronariopata
CONT	–	Criança com Peso Normal
COB	–	Criança com Sobrepeso
APN	–	Adulto com Insufiência Respiratória
NOR	–	Adulto com Ritmo Sinusal Normal
CHF	–	Adulto com Falha Congestiva Cardíaca
S810i	–	Modelo do Monitor Cardíaco Polar
RS800	–	Modelo do Monitor Cardíaco Polar
ANOVA	–	Teste da Variância

SUMÁRIO

Pág.

1 INTRODUÇÃO

Na literatura, há diversos trabalhos que estudam a capacidade de adaptação de um indivíduo em interação com o ambiente. Nesses estudos são analisadas séries temporais de ritmo cardíaco ou frequência cardíaca[1] (PAGANI, 2000; ACHARYA et al., 2006; VANDERLEI et al., 2009; MAZON et al., 2013) para compreensão das variações realizadas pelo coração ao longo do tempo em suas interações com o ambiente.

Sabe-se que todo o sistema cardíaco está ligado diretamente ao funcionamento adequado do sistema nervoso. Assim acredita-se que através da análise cardíaca pode-se predizer o estado do sistema nervoso do organismo (PAGANI, 2000). Em particular, o coração pode ser entendido como um sistema dinâmico e não estacionário, uma vez que está em constante mudança para se adaptar às variações ambientais, emocionais e físicas do indivíduo.

Dentre as várias formas de analisar as alterações do coração, pode-se citar o eletrocardiograma (VANDERLEI et al., 2009) que é uma técnica não invasiva e que capta a *Variabilidade da Frequência Cardíaca*[2] (VFC). Segundo Task Force of the European Society of Cardiology the North American Society of Pacing Electrophysiology (1996), a análise da VFC representa uma significativa contribuição para a compreensão da relação entre a dinâmica do coração e o sistema nervoso autônomo. Isso se deve à popularização da VFC como uma medida de fácil captação.

Podemos também fazer comparações entre grupos de indivíduos que podem apresentar sistemas semelhantes, por exemplo, comparar grupos de séries temporais de ritmo cardíaco de recém-nascidos prematuros e recém-nascidos normais. (SELIG et al., 2011). Espera-se que a VFC desses dois grupos seja equivalente em alguns pontos e diferente em outros. Usualmente, para análise são usados diversos métodos: no domínio temporal (RAMÍREZ-ROJAS; FLORES-MÁRQUEZ, 2013), no domínio da frequência (OLIVEIRA et al., 2012) e dinâmica não linear (TRULLA et al., 1996; SUNKARIA, 2011).

Os métodos no domínio do tempo estão tradicionalmente relacionados a uma análise linear do comportamento temporal. Esta seria obtida a partir da análise da série

[1] Séries temporais de ritmo cardíaco ou frequência cardíaca estão relacionadas à capacidade do sistema cardíaco e sistema nervoso do indivíduo em se adaptar a diferentes variações do ambiente (VANDERLEI et al., 2009).

[2] A Variabilidade da Frequência Cardíaca é resultado da adaptação cardíaca as diversas atividades demandadas pelo organismo (Task Force of the European Society of Cardiology the North American Society of Pacing Electrophysiology, 1996).

temporal empregando-se uma série transformada gerada a partir das diferenças adjacentes entre os pontos consecutivos das séries temporais e outros parâmetros estatísticos como média e desvio padrão (Task Force of the European Society of Cardiology the North American Society of Pacing Electrophysiology, 1996). Nota-se que os métodos lineares no domínio temporal podem não caracterizar de forma apropriada o sistema sob análise, uma vez que sabidamente o ritmo cardíaco apresenta uma dinâmica não linear.

Os métodos no domínio da frequência também extensivamente usados, geralmente baseados na tranformada de Fourier, sabidamente apresentam limitações na análise de sinais (ACHARYA et al., 2006) e fornecem uma interpretação das estruturas regulares do conjunto de dados. Tanto os métodos no domínio temporal e no domínio da frequência analisam a dinâmica intrínsica dos sistemas, cuja caracterização seria precisa se fossem governados por um paradigma linear (KANTZ; SCHREIBER, 2004), o que não é o caso.

Uma alternativa eventualmente mais realista e apropriada diante do fato de se estar considerando um sistema reconhecidamente não linear é o emprego dos métodos de análise oriundos da teoria dos sistemas dinâmicos não lineares (BRENNAN et al., 2001; GLASS, 2009; GONCALVEZ et al., 2013). Com esse enfoque, tem-se ferramentas em tese mais apropriadas e capazes de detectar todo o comportamento multifacetado que está presente na dinâmica do coração. Entretanto, cada método pode possuir uma sensibilidade diferente ao comparar dois sistemas semelhantes, que precisa ser adequadamente avaliado, considerando-se o problema em questão.

Muitos trabalhos usam diversas ferramentas para análise da VFC originárias de métodos empregados na análise de sistemas dinâmicos não lineares, como são os casos da dimensão de correlação, expoentes de Lyapunov e entropia de Shannon (POOL, 1989; FREITAS et al., 2009; KAREMAKER; BERECKI-GISOLF, 2009).

Para caracterizar dinâmicas muito próximas a partir de séries temporais não lineares, nesse trabalho propomos o uso associado de diversos métodos de duas áreas do conhecimento que se complementam: os métodos de sistemas dinâmicos (não lineares) (BAKER; GOLLUB, 1998; KANTZ; SCHREIBER, 2004) e as técnicas de mineração de dados (WITTEN; FRANK, 2005). Esses métodos são usados basicamente em duas etapas:

- obtenção de índices paramétricos dos conjuntos de séries temporais é realizada pelos métodos de sistemas dinâmicos;

- os índices paramétricos obtidos com os métodos de sistemas dinâmicos são usados como padrões de entrada para as técnicas de mineração de dados tranformando em informações sobre os sinais analisados.

Os métodos de sistemas dinâmicos usados são: o modelo autoregressivo (De Carvalho et al., 2002), o gráfico de Poincaré (PISKORSKI; GUZIK, 2007), a medida da tendência central (COHEN et al., 1996), dinâmica simbólica (GUZZETTI et al., 2005), a medida de complexidade (PIQUEIRA; MATTOS, 2011) e as medidas de quantificação da recorrência (ECKMANN et al., 1987; MARWAN et al., 2002; MARWAN et al., 2007). Esses métodos fornecem medidas das séries temporais que serão usadas como padrões de entradas das técnicas de mineração de dados. São usadas duas técnicas, o classificador J48 (SAFAVIAN; LANDGREBE, 1991) que gera uma árvore de decisão, e máquinas de vetores de suporte (CORTES; VAPNIK, 1995). Cada método que compõe a metodologia contribui na compreensão da comportamento do coração usando a caracterização dada pelos métodos de mineração de dados. Com esse objetivo, algumas aplicações dessas técnicas são apresentadas a seguir.

O modelo autoregressivo, também conhecido como modelo de máxima entropia em aplicações físicas (De Carvalho et al., 2002), permite a construção de um modelo linear (JUNG; PARK, 2001; AGUIRRE, 2007). Esse método permite verificarmos o quanto as séries temporais estudadas se distanciam do modelo criado. Podemos dizer que esse modelo é utilizado como um "preditor" do comportamento não linear das séries temporais de VFC. O gráfico de Poincaré ou mapa de primeiro retorno é considerado uma ferramenta de análise visual que permite obter parâmetros que caracterizam aspectos importantes relacionados à VFC de curta e longa duração (PISKORSKI; GUZIK, 2007). O uso desse método é amplamente difundido e auxilia na análise da influência do sistema nervoso simpático conforme descrito em Woo et al. (1994). Esse método é usado na compreensão dos fenômenos fisiológicos e biológicos ocorridos durante treinamento físico (MOUROT et al., 2004), em pacientes com insuficiência cardíaca congestiva (ISLER, 2007), durante a prática da meditação (GOSHVARPOUR et al., 2011), na detecção de condições de estresse no indivíduo (MELILLO et al., 2011) e na detecção de alterações na VFC após estimulação acústica em adultos saudáveis (ROY et al., 2012).

A medida da tendência central apresentada por Cohen et al. (1996) é muito usada em modelagem de sistemas biológicos, hemodinâmica e análise da variabilidade da frequência cardíaca (JEONG et al., 2002). Esse método pode melhorar a capacidade de diagnóstico de pacientes com suspeita clínica de apnéia obstrutiva do sono, usando

sinais de oximetria (quantificação de oxigênio no sangue) (ALVAREZ et al., 2007). Outro estudo mostra que a medida da tendência central, combinada com outros métodos auxilia na diferenciação entre VFC normal e anormal (JOVIC; BOGUNOVIC, 2011).

A dinâmica simbólica é um método que transforma uma série temporal em um série de símbolos. Esse método pode ser aplicado para quantificar a prevalência da modulação cardíaca simpática e parassimpática elucidando os mecanismos que ocorrem durante os curtos períodos que antecedem eventos cardíacos agudos (GUZZETTI et al., 2005). Outro estudo usa a dinâmica simbólica para analisar a VFC em pacientes com doença arterial periférica submetidos a uma única sessão de exercício físico (LIMA et al., 2013). Já a medida de complexidade é um método para análise de sinais obtida pelo produto da ordem e desordem (PIQUEIRA; MATTOS, 2011), sendo sua primeira aplicabilidade para a análise da VFC realizada neste trabalho.

As medidas de quantificação de recorrência (RQA) também são usadas para estimar características dos sistemas dinâmicos (ECKMANN et al., 1987; MARWAN et al., 2002; MARWAN et al., 2007). O método RQA pode ser usado para diferenciar a dinâmica de sinais de pacientes coronariopatas e de sujeitos normais (GUO et al., 2012). Segundo (GUO et al., 2012) pacientes coronariopatas têm maior regularidade, determinismo, estabilidade e menor variabilidade que sujeitos normais. Outro estudo usa RQA para investigar o efeito do desafio ortostático[3] na VFC e da pressão sanguínea, mostrando que esse método é capaz de diferenciar o indivíduo em supino (deitado) e posição ereta pelas adaptações do sistema cardiovascular (JAVORKA et al., 2009).

A partir desse conjunto de métodos oriundos da análise de sistemas dinâmicos, é obtido um conjunto de índices das séries temporais de VFC. Esses índices são usadas como padrões de entrada nos classificadores de árvore de decisão e máquinas de vetores de suporte. Tais classificadores também possuem aplicabilidade na análise da dinâmica do coração, conforme descrito em (JOVIC; BOGUNOVIC, 2011; GUO et al., 2012; YU; LEE, 2012).

É importante salientar que neste trabalho esses métodos são aplicados exaustivamente em diferentes comparações entre grupos de séries temporais com comportamentos semelhantes. Esse procedimento é adotado para estabelecermos o conjunto de métodos adequados para cada grupo de sinais não lineares estudados. Outro aspecto relevante da aplicação dessa metodologia nos sinais de VFC é o volume de

[3]Desafio ortostático está relacionado ao indivíduo estar na posição ereta.

dados analisados para esse estudo de caso. O conjunto de dados é constituído de 514 séries temporais de ritmo cardíaco.

Inicialmente, para analisar o conjunto de séries temporais de ritmo cardíaco ou séries temporais experimentais de uma forma geral, é necessário que estas séries passem por um processo de filtragem com a finalidade de serem excluídas informações que não fazem parte da série em si (KARLSSON et al., 2012). No caso de análise sequencial de intervalos entre os batimentos cardíacos é imprescindível a exclusão de artefatos (por exemplo, duplo reconhecimento), intervalos não oriundos do nó sinoatrial do coração (complexos ventriculares prematuros - VPC) (WESSEL et al., 2000), e interferências espúrias como as devidas a tremores musculares, má colocação de eletrodos, efeito de equipamentos eletrônicos estranhos no ambiente de captação do sinal, etc (LOGIER et al., 2004; KEENAN; GROSSMAN, 2006).

Neste trabalho usamos uma filtragem adaptativa automática que permite uma pré-análise de um grande volume de dados. O principal objetivo em usar uma filtragem automática adaptativa, neste contexto, é verificar o desempenho do filtro comparando os resultados com os obtidos com o processo de filtragem tradicionalmente usado pelo especialista em análise de séries temporais de ritmo cardíaco.

Verificamos também que o conjunto de métodos apresentado para discriminação das séries temporais de intervalos RR pode ser aplicado em outros estudos envolvendo grupos de sinais não lineares. De fato, acreditamos que temos uma estratégia sensível o suficiente que nos permite discriminar conjuntos que apresentam dinâmicas muito próximas. Visando corroborar essa nossa afirmação, a metodologia é aplicada em um segundo estudo de caso, relacionado à compreensão da alta atmosfera e para tal, podemos usar séries temporais de vento neutro[4] (ANDRIOLI, 2012; FRITTS et al., 2012).

Essas séries temporais de vento neutro são captadas dos radares meteóricos de duas localidades distintas: na cidade de Cachoeira Paulista no Estado de São Paulo e na Estação Comandante Ferraz na Base Brasileira na Antártica. Antevê-se que esses dois conjuntos de séries possuam uma dinâmica distinta entre si, pois essas séries temporais de vento neutro podem apresentar significativa variabilidade conforme a estação do ano e latitude, devido a variações sazonais em suas fontes e ambientes de propagação (FRITTS et al., 2012). Ao todo esse conjunto é constituído de 47 séries temporais e que são classificadas em dois grupos conforme a localização do radar

[4]Séries temporais de vento neutro são detectadas na alta atmosfera e estão relacionadas à detecção dos meteoros que entram na atmosfera.

meteórico.

1.1 Objetivo

Este trabalho tem por objetivo principal discriminar conjuntos de sinais não lineares pelo uso associado de métodos de sistemas dinâmicos e técnicas de mineração de dados. São usadas 514 séries temporais de frequência cardíaca que foram pré-classificadas por um médico especialista em nove grupos: recém-nascidos prematuros, recém-nascidos normais, crianças com peso normal, crianças com sobrepeso, jovens adultos saudáveis, adultos em dieta de baixa calorias, adultos coronariopatas, adultos com insuficiência respiratória e adultos com falha cardíaca congestiva.

Inicialmente todas as séries temporais são pré-processadas usando uma filtragem que remove pontos considerados, segundo critérios estabelecidos pelo filtro, artefatos ou interferências no sinal analisado. Dado um grande volume de dados, esse processo de filtragem é automatizado para facilitar o procedimento tradicional utilizado por um especialista em análise das séries analisadas. A comparação entre os processos de filtragem: adaptativo e o tradicional ou convencional é realizada (após os ajustes dos parâmetros do filtro) para verificar a significância estatística em usarmos um método automatizado, facilitando assim o pré-processamento dos sinais. Os parâmetros do filtro são escolhidos de maneira adequada de acordo com os conjuntos de séries estudadas[5].

Avaliamos também as técnicas originárias da teoria de sistemas dinâmicos não lineares e usados para obtenção dos índices paramétricos. A eficácia do enfoque é dada pela diferenciação dos valores entre os sistemas estudados. Entretanto, sabemos que nem todos os métodos são sempre adequados para diferenciar todos os conjuntos de dados. Portanto, estabelecemos o grupo específico de medidas para cada comparação dos conjuntos de séries temporais estudadas.

Nesse contexto, mostramos as medidas específicas capazes de diferenciar os conjuntos de séries temporais de frequência cardíaca e para o conjunto de séries temporais de vento neutro.

[5] O ajuste dos parâmetros do filtro e a comparação entre os dois tipos de filtragem são abordados no Capítulo 3.

2 METODOLOGIA

Esse capítulo apresenta a metodologia elaborada para análise de sinais não lineares. É apresentada inicialmente uma definição da metodologia usada, após uma definição de sistema dinâmico e suas classificações, abordando as etapas de extração das medidas (técnicas de dinâmica linear e não linear) e a caracterização usando as técnicas de mineração de dados.

Em aspectos gerais essas técnicas são usadas para extrair valores das séries temporais analisadas e que serão utilizadas como informações de entrada nos classificadores.

2.1 Definição

A metodologia proposta nesse trabalho envolve a junção de duas diferentes áreas, métodos de sistemas dinâmicos e mineração de dados, na análise de séries temporais com comportamento não linear. Esta junção permite compreender a relação existente entre as medidas que são extraídas das séries temporais e o contexto no sistema no qual estão inseridas.

A metodologia proposta envolve quatro etapas. A primeira etapa refere-se ao conjunto de séries. Neste trabalho foram escolhidos conjuntos de séries temporais que sabidamente possuem um comportamento não linear, tal como conjunto de séries temporais de intervalos RR[1].

A segunda etapa é o pré-processamento das séries temporais. Esse pré-processamento é realizado de acordo com o tipo de série analisada. Para o conjunto de séries temporais de intervalos RR o procedimento é a filtragem, que remove os artefatos. Já para o conjunto de séries temporais de vento zonal o procedimento é a interpolação das séries. Detalhes sobre cada pré-processamento utilizado é abordado separadamente em cada estudo de caso.

A terceira etapa da metodologia envolve a obtenção dos valores (índices paramétricos) dos conjuntos de séries temporais. Para tal, são usados métodos de dinâmica linear e não linear. Essas medidas servem como um conjunto de entrada/informações para os métodos de mineração de dados. E por fim, na quarta etapa os métodos de mineração de dados, que são os classificadores de árvore de decisão e máquinas de vetores de suporte, são usados para caracterizar a dinâmica em que essas séries temporais estão inseridas.

[1] Séries temporais de intervalos RR ou tacogramas são séries que quantificam a variação das ondas R presentes em um eletrocardiograma (detalhes no Capítulo 3).

2.1.1 Definição de sistemas dinâmicos

O termo sistema é usado em diversas áreas do conhecimento, como na informática, administração, medicina e biologia. Apesar da diversidade de áreas, este termo converge para um mesmo sentido, o de agrupar e combinar diferentes elementos.

Segundo Monteiro (2006) um sistema é "um conjunto de objetos agrupados por alguma interação ou interdependência, de modo que existam relações de causa e efeito nos fenômenos que ocorrem com os elementos desse conjunto". Alguns exemplos de sistemas podem ser o planeta Júpiter e seus satélites naturais, os órgãos do corpo humano, o ecossistema de uma floresta, equações que descrevem um determinado comportamento, etc. Um sistema pode ser considerado *dinâmico* quando esse possui grandezas relacionadas aos objetos constituintes que variam no tempo.

Os sistemas dinâmicos são classificados de acordo com certas características, tais como evolução temporal ou possibilidade de previsibilidade. Determinar a dinâmica da evolução temporal das grandezas de um sistema é importante pois possibilita estudar um sistema que ainda não existe fisicamente. Por exemplo, o comportamento de um satélite artificial a ser construído, compreender características de um sistema já existente, entender questões da Biologia ligadas a organismos vivos e sua evolução ou ainda simular a dinâmica de experimentos muito caros ou perigosos, como em campanhas de vacinação (AGUIRRE, 2007).

Para estudar e caracterizar a dinâmica dos sistemas através do uso de séries temporais, podemos aplicar técnicas baseadas em dinâmicas linear e não linear. Nesse estudo são usadas diversas técnicas com a finalidade de estabelecer quais caracterizam de forma apropriada as séries temporais estudadas fornecendo mais informações a nossos propósitos sobre sua dinâmica.

2.2 Obtenção dos índices paramétricos

2.2.1 Método autoregressivo

O método autoregressivo (AR), também conhecido como modelo de máxima entropia em aplicações físicas (De Carvalho et al., 2002), permite a construção de um modelo que se aproxima da série analisada (JUNG; PARK, 2001). Considerando uma série temporal $x = x_t, x_{t+1}, x_{t+2}, ..., x_{t+n}$, o modelo autoregressivo pode ser definido como:

$$x_t = \sum_{i=1}^{N} a_i x_{t-i} + \xi_t \tag{2.1}$$

onde x_t é ponto da série sob investigação no tempo t, a_i para $i = 1, ..., N$ é o coeficiente autoregressivo, N (nesse estudo $N = 1$) é a ordem do filtro, que é geralmente menor que o comprimento da série e ξ_t é o ruído ou resíduo (geralmente ruído branco gaussiano). Para a resolução da Equação 2.1, ou seja, estimar seus parâmetros a partir de uma série temporal, pode ser usado o método dos mínimos quadrados (AGUIRRE, 2007).

Uma alternativa à utilização do modelo AR em séries temporais sabidamente com uma dinâmica não linear é usar uma "janela móvel de tempo", que gera ao longo do tempo uma série temporal de tamanho menor, para o qual os parâmetros do modelo AR são calculados. Ou seja, a cada determinado intervalo de tempo estima-se a_i. Após estimar o conjunto de coeficientes obtém-se a nova série temporal x', fornecida a partir do modelo AR calculando-se assim o erro quadrático médio (EQM) entre as duas séries (original e estimada):

$$EQM = \frac{1}{n}\sum_{t=1}^{n}(x'_t - x_t)^2 \tag{2.2}$$

onde n é o tamanho de cada série, x'_t é a série estimada no instante de tempo t e x_t é a série original. Quanto maior o valor de EQM, mais diferente é a série estimada da série original mostrando que apenas os coeficientes lineares associados às janelas podem não se mostrarem adequados para modelar o sistema. Ou seja, se o ruído presente no sistema estiver compatível com as hipóteses do modelo AR, o EQM da série original está relacionado a um comportamento não linear ou, eventualmente, é necessário calcular coeficientes de maior ordem.

2.2.2 Gráfico de Poincaré

O gráfico de Poincaré, denominado também como mapa de primeiro retorno, pode ser definido como um método de análise onde cada valor da série temporal analisada é graficado em função do valor anterior. Por exemplo, dada a série temporal $x = x_t, x_{t+1}, x_{t+2}, ..., x_{t+n}$, no gráfico de Poincaré estarão os pontos (x_1,x_2), após (x_2,x_3) e assim sucessivamente, ou seja, (x_i,x_{i+1}).

Em especial, essa técnica é muito usada na análise de séries temporais de intervalos RR relacionadas ao sistema cardíaco (BRENNAN et al., 2001; MOUROT et al., 2004; ISLER, 2007; VANDERLEI et al., 2009; KARMAKAR et al., 2009; GOSHVARPOUR et al., 2011; PETKOVIC; COJBASIC, 2012). Nessas séries de intervalos RR, cada valor da série (intervalo RR) é a duração de um batimento cardíaco ou, mais precisamente,

a duração entre duas ondas R (ver Seção 3.1).

Esse método pode ser considerado uma ferramenta visual capaz de possibilitar que se analisem características consideradas relevantes associadas a uma série de intervalos RR derivada de um eletrocardiograma, além de permitir que se obtenham parâmetros quantitativos que permitem caracterizar aspectos importantes relacionadas à VFC longa e de curta duração (PISKORSKI; GUZIK, 2007).

Do gráfico de Poincaré podem ser extraídos alguns parâmetros que fornecem informações sobre a dinâmica da série temporal analisada. Dentre eles, os seguintes parâmetros: SD1, SD2 e SD1/SD2 (ver Figura 2.1). X_2 corresponde à linha identidade no gráfico. No gráfico os pontos representam as coordenadas x_i,y_i para $i = 1, ..., n + 1$, ou seja RR_i,RR_{i+1}. Sendo RR_i o $i-ésimo$ intervalo RR e n é o número de pontos no gráfico de Poincaré (que é um intervalo RR menor que o comprimento da série temporal). X_1 é definido como $X_1 = -x + 2 * \overline{x}$, onde $\overline{x}$ é a média dos pontos em x. O eixo x, denominado $RR_n(s)$, no gráfico de Poincaré corresponde aos intervalos RR_1 a RR_n. O centróide $(RR_{iC}, RR_{(i+1)C})$ é definido como:

$$\begin{aligned} RR_{iC} &= \overline{x} \\ RR_{(i+1)C} &= \overline{y} \end{aligned} \tag{2.3}$$

onde $\overline{y}$ é a média dos pontos em y. O eixo y, denominado $RR_{n+1}(s)$, no gráfico de Poincaré corresponde aos intervalos RR_2 a RR_{n+1}, sendo $n + 1$ o tamanho total da série de intervalos RR.

O parâmetro SD1 (está relacionado com a variabilidade de curta duração), obtido ao longo da reta perpendicular (X_1) à linha identidade, é definido conforme descrito em (BRENNAN et al., 2001):

$$\begin{aligned} SD1^2 &= Var(X_1) \\ &= Var\left(\frac{1}{\sqrt{2}}RR_n - \frac{1}{\sqrt{2}}RR_{n+1}\right) \\ &= \frac{1}{2}SDSD^2 \end{aligned} \tag{2.4}$$

onde SDSD corresponde ao desvio padrão das sucessivas diferenças dos intervalos RR.

Segundo Brennan et al. (2001) o parâmetro SD2 (descreve a variabilidade de longa

duração), desvio padrão da projeção para linha da identidade (X_2), é definido como:

$$SD2^2 = 2SDRR^2 - \frac{1}{2}SDSD^2 \tag{2.5}$$

sendo SDRR corresponde ao desvio padrão de todos os intervalos RR normais gravados em um intervalo de tempo. Pode ser definido como:

$$SDRR = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(RR_i - \overline{RR})^2} \tag{2.6}$$

onde $\overline{RR}$ denota a média dos intervalos RR. Os parâmetros do gráfico de Poincaré são ilustrados na Figura 2.1.

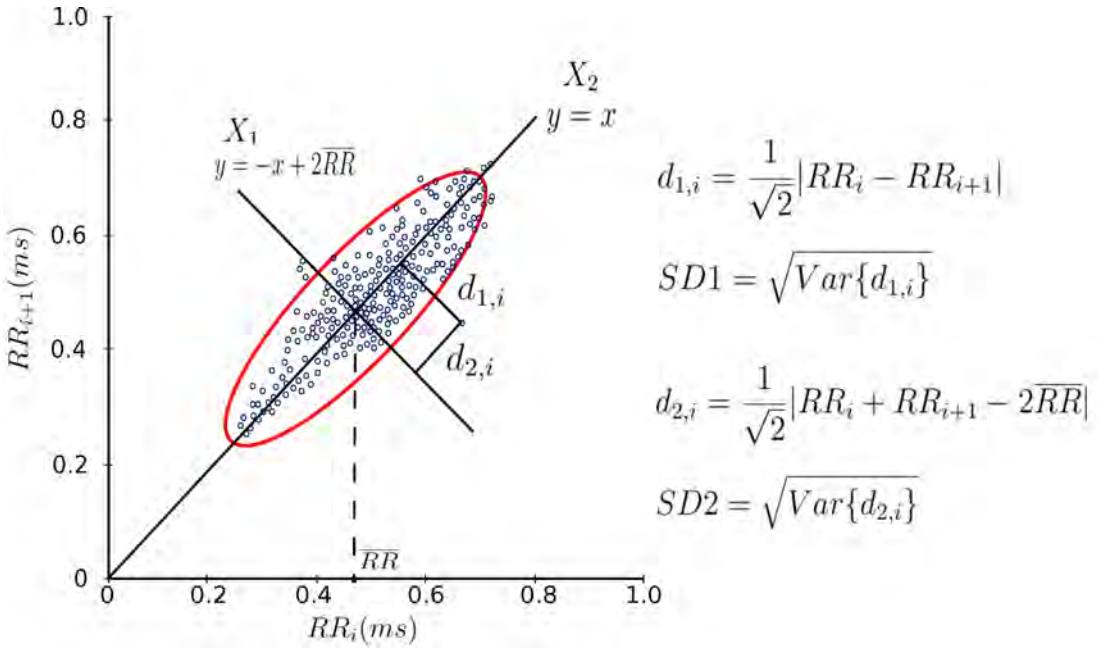

Figura 2.1 - Esquema ilustrativo dos parâmetros do gráfico de Poincaré.

O parâmetro SD1/SD2, obtido a partir do gráfico de Poincaré, é denominado *razão de dispersão*. Para os sinais cardíacos, essa medida representa a razão entre a variação do intervalo curto e a variação do intervalo longo. Em outras palavras, SD1/SD2 é a razão de dispersão da diferença entre os intervalos de tempo de pulsos adjacentes pela dispersão global dos intervalos de tempo dos batimentos.

2.2.3 Medida da Tendência Central

Representada pelo gráfico de diferenças de segunda ordem, a medida da Tendência Central (CTM) é muito usada em modelagem de sistemas biológicos, hemodinâmica e análise da variabilidade da frequência cardíaca (JEONG et al., 2002).

Esses gráficos de espalhamento fornecem um diagnóstico característico via inspeção visual do diagrama (ALVAREZ et al., 2007). Os pontos representativos de uma série temporal com maior variabilidade nos sucessivos intervalos de tempo apresentam-se mais dispersos no gráfico do em que séries temporais com menor variabilidade nos intervalos de tempo. Para ilustrar esse comportamento, seguem os exemplos abaixo.

Exemplo: série com valores constantes

Considerando a Figura 2.2A a série temporal, onde $i = 1, ..., 15$ representa a sucessão dos pontos da série e $x_i = 1$ o valor repetitivo para todos os pontos sucessivos da série. Tem-se que o gráfico de espalhamento de segunda ordem para os pontos $(x_{i+2} - x_{i+1}, x_{i+1} - x_i)$ não variam e apresentam-se concentrados na origem (0,0), evidenciando sobreposição dos pontos, conforme mostrado na Figura 2.2B. Esse aspecto do gráfico evidencia que a série não apresenta diferenças sucessivas entre os pontos.

Exemplo: série temporal do mapa logístico

A Figura 2.3A mostra uma série temporal do mapa logístico[2], considerando $\rho = 3.5$ e $x_1 = 0,5$, quando ainda não há presença de caos no sistema, apresentando periodicidade. No gráfico de espalhamento da Figura 2.3B verifica-se a presença de quatro pontos (mostrando sobreposição de pontos), evidenciando a baixa variabilidade da série em relação às diferenças sucessivas.

Exemplo: série temporal aleatória

A Figura 2.4 mostra o gráfico de espalhamento para uma série aleatória. Observe que tal série apresenta um padrão variável de distribuição de pontos, e tal padrão manifesta-se no gráfico de espalhamento. Na Figura 2.4B verifica-se que os pontos $(x_{i+2} - x_{i+1}, x_{i+1} - x_i)$ apresentam-se dispersos.

[2]O mapa logístico é definido como: $x_{n+1} = \rho x_n(1 - x_n)$, sendo ρ um parâmetro de taxa de crescimento quando trata-se do modelo de crescimento populacional. Esse mapa discreto pode apresentar comportamento caótico quando ρ oscila próximo de 4.

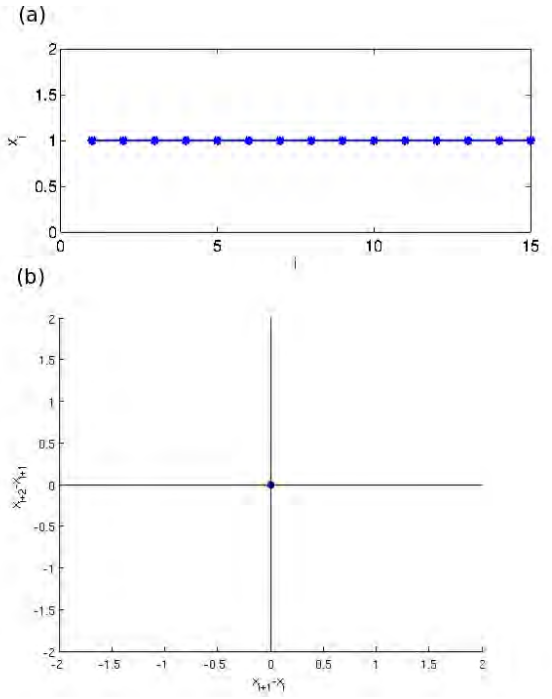

Figura 2.2 - a) Série com 15 valores constantes. b) Gráfico de espalhamento de segunda ordem para a série com valores constantes. Apresenta um único ponto na origem (0,0) evidenciando que não há diferenças sucessivas entre os pontos da série.

A CTM quantifica o grau de variabilidade da série analisada no gráfico de espalhamento e é calculada definindo-se uma região circular de raio ρ (escolhido conforme a característica do dado) em torno da origem, contando-se o número de pontos no interior do círculo e dividindo-se o resultado pelo total de pontos. Dado N pontos na série temporal de intervalos RR, $N-2$ é a quantidade de pontos no gráfico. A CTM é calculada como (COHEN et al., 1996):

$$CTM = \frac{\sum_{i=1}^{N-2} \delta(d_i)}{N-2} \tag{2.7}$$

onde

$$\delta(d_i) = \begin{cases} 1 \text{ se } [(x(i+2)-x(i+1))^2 + (x(i+1)-x(i))^2]^{1/2} < \rho \\ 0 \quad \text{caso contrário} \end{cases} \tag{2.8}$$

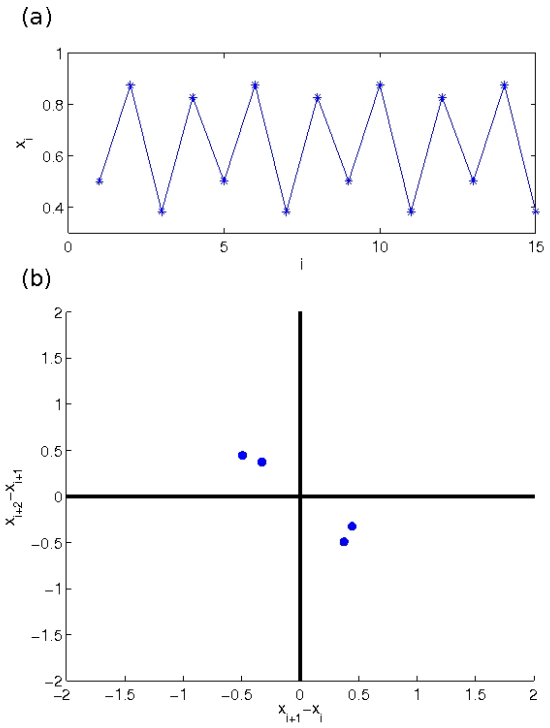

Figura 2.3 - a) Série do mapa logístico para $\rho = 3.5$ e $x_1 = 0,5$ com 15 pontos. b) Gráfico de espalhamento de segunda ordem para a série, apresentando apenas quatro pontos, evidenciando a baixa variabilidade nas diferenças sucessivas.

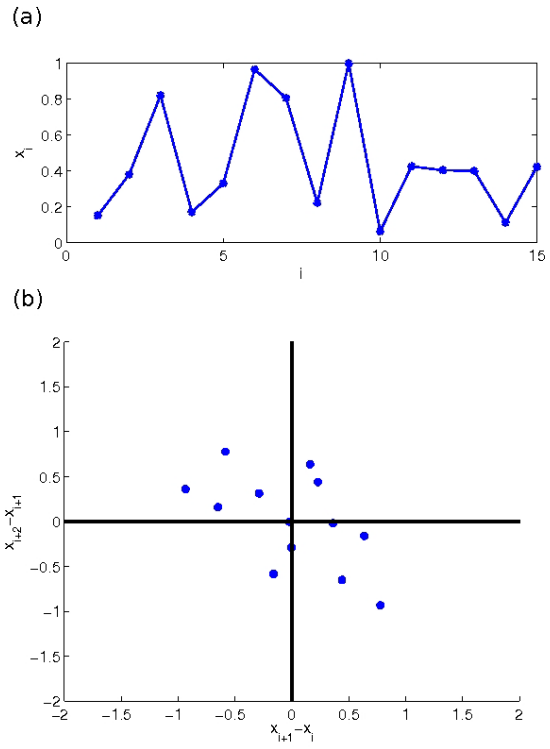

Figura 2.4 - a) Série aleatória com 15 pontos. b) Gráfico de espalhamento para a série apresentando pontos dispersos, evidenciando maior variabilidade em relação às diferenças sucessivas.

2.2.4 Dinâmica simbólica

Dinâmica simbólica contiste aqui na transformação de uma série temporal $\Delta RR = (\delta RR_1, \delta RR_2, ..., \delta RR_N)$, onde cada $\delta RR_i = x_{i+1} - x_i$ sendo $i = 1, 2, ..., N-1$, em uma série de símbolos $S = \{s_1, s_2, ..., s_N\}$ pertencente a um alfabeto finito (SMALE, 1967).

Nesse estudo são usados três símbolos (0, 1 e 2) da seguinte maneira:

$$s_i = \begin{cases} 0 \text{ se } |\delta RR_i| \leq \tau \\ 1 \text{ se } \delta RR_i > \tau \\ 2 \text{ se } \delta RR_i < -\tau \end{cases} \tag{2.9}$$

Dada a sequência S de símbolos, temos a sequência W de palavras, $W = (w_1, w_2, w_3, ..., w_{N-(l-1)})$, onde l é o tamanho da palavra usada e $w_i = (s_i, s_{i+1}, s_{i+2}, ..., s_{l+i-1})$.

Baseado na frequência de palavras de tamanho l na série S, por exemplo, palavra com tamanho $l = 5$ fornece $3^l = 243$ diferentes palavras, sendo possível calcular para o conjunto produzido pela dinâmica do sistema vários quantificadores de caracterização da dinâmica, como, por exemplo, a entropia de Shannon:

$$H = -\sum_{k=1}^{3^l} w_k \ln w_k \tag{2.10}$$

para probabilidade da palavra w_k.

2.2.5 Medida de complexidade

A medida de complexidade ($\Gamma_{\alpha\beta}$) é um produto entre desordem (Δ) e ordem ($1-\Delta$) associada à entropia de informação (H) (PIQUEIRA; MATTOS, 2011). Logo, a medida de complexidade pode ser:

$$\Gamma_{\alpha\beta} = (1-\Delta)^{\alpha}(\Delta)^{\beta} \quad \text{onde} \tag{2.11}$$

$$\Delta = \frac{H}{H_{max}} \tag{2.12}$$

α e β são constantes positivas, escolhidas de acordo com o objeto de estudo. Sendo que α dá peso ao termo ordem e β para o termo desordem. A entropia (H) está

definida na Equação 2.10.

Se $\alpha = 0$, somente o termo desordem é considerado. Considerando somente o termo ordem ($\beta = 0$), um parâmetro de organização surge para expressar complexidade. Como estas situações são triviais, casos com $\alpha > 0$ e $\beta > 0$ requerem uma análise mais extensiva.

Nesse trabalho, fixamos o valor de $\alpha = 1$ e foram testados alguns valores de $\beta = 0,25$, $\beta = 0,5$ e $\beta = 1$.

2.2.6 Análise de Recorrência

O termo recorrência está associado a Henri Poincaré ($1854 - 1912$), quando em 1890 publicou o "teorema da recorrência de Poincaré" (POINCARÉ, 1892). Esse teorema garante que, para um conjunto grande de sistemas dinâmicos, as trajetórias retornam infinitas vezes, arbitrariamente próximas a quase todos os pontos iniciais, formando um conjunto infinito de instantes de retorno[3]. Na prática, em alguns sistemas como os caóticos torna-se impossível encontrar recorrência total, isto é, o estado de um sistema caótico não pode recorrer exatamente ao seu estado inicial, tornando-se obrigatório o uso de uma vizinhança $m-$dimensional (BAKER; GOLLUB, 1998) (ver Figura 2.5).

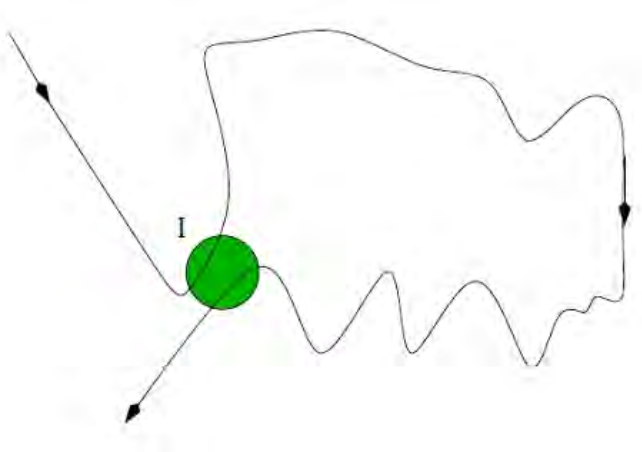

Figura 2.5 - Ilustração da recorrência de Poincaré em um conjunto I qualquer com espaço bidimensional arbitrário.
Fonte: Souza (2008)

[3]Considera-se instante de retorno ou mapa de retorno quando a trajetória de um ponto retorna a condição inicial.

Baseado na idéia de Poincaré, Eckmann et al. (1987) introduziram uma ferramenta conhecida como gráfico de recorrência, que serve para visualizar a dinâmica de sistemas recorrentes.

O gráfico de recorrência de uma série temporal $x(t_N)$ de N pontos pode ser definido como uma matriz $N \times N$ preenchida por pontos brancos e pretos. O ponto preto é chamado de recorrente, é colocado na matriz de recorrência com coordenadas (i,j) somente se a distância $\epsilon(i,j)$ no instante $n = i$ e $n = j$ (entre o estado corrente do sistema e o estado a ser comparado) for menor que uma distância definida (raio) ϵ_0, fixado no centro do estado corrente.

A representação bidimensional da matriz gerada no gráfico de recorrência pode ser dada através da relação:

$$R_{i,j} = H(\epsilon_0 - ||\xi_i - \xi_j||), \quad \xi_i \in R^m, \quad i,j = 1,...,N \tag{2.13}$$

onde N é o número de estados ξ_i considerados, ϵ_0 é o raio de vizinhança (*threshold*),$||.||$ é a norma euclidiana, $H(x)$ é a função de Heaviside e m a dimensão de imersão (maiores detalhes ver (KANTZ; SCHREIBER, 2004)).

Se $R_{i,j} = 1$, o estado é dito recorrente e, como conseqüência, um ponto preto é marcado no gráfico de recorrência. Caso $R_{i,j} = 0$, o estado é não recorrente e um ponto branco é marcado no gráfico.

O gráfico de Recorrência apresenta, conforme a série analisada, diferentes padrões visuais que podem fornecer informações sobre o dado analisado. Eckmann et al. (1987) distinguiram (arbitrariamente) através da análise de gráficos de Recorrência entre os padrões de tipologia de larga escala e os padrões de textura de pequena escala. Enquanto a tipologia de grande escala fornece uma visão global do sistema analisado, a textura de pequenas escalas fornece uma análise mais particular do sistema (SOUZA, 2008). Segundo Marwan et al. (2007) os padrões de larga escala nos gráficos de Recorrência (tipologia) podem ser classificados como homogêneo, periódico, deriva (*drift*) e descontínuo (*disrupted*):

- Homogêneo: típico de sistemas estacionários, por exemplo, um gráfico de recorrência de uma série temporal aleatória (ver Figura 2.6A);
- Periódico: típico de sistemas periódicos ou quasi-periódicos, apresentam diagonal orientada com estruturas recorrentes (linhas diagonais e estruturas

"tabuleiro de damas") (ver Figura 2.6B);

- Deriva: típico de sistemas com variação de parâmetros lenta, isto é, sistemas não estacionários (ver Figura 2.6C);
- Descontínuo: típico de gráficos de recorrência com mudanças na dinâmica como eventos extremos causando áreas ou bandas brancas (ver Figura 2.6D).

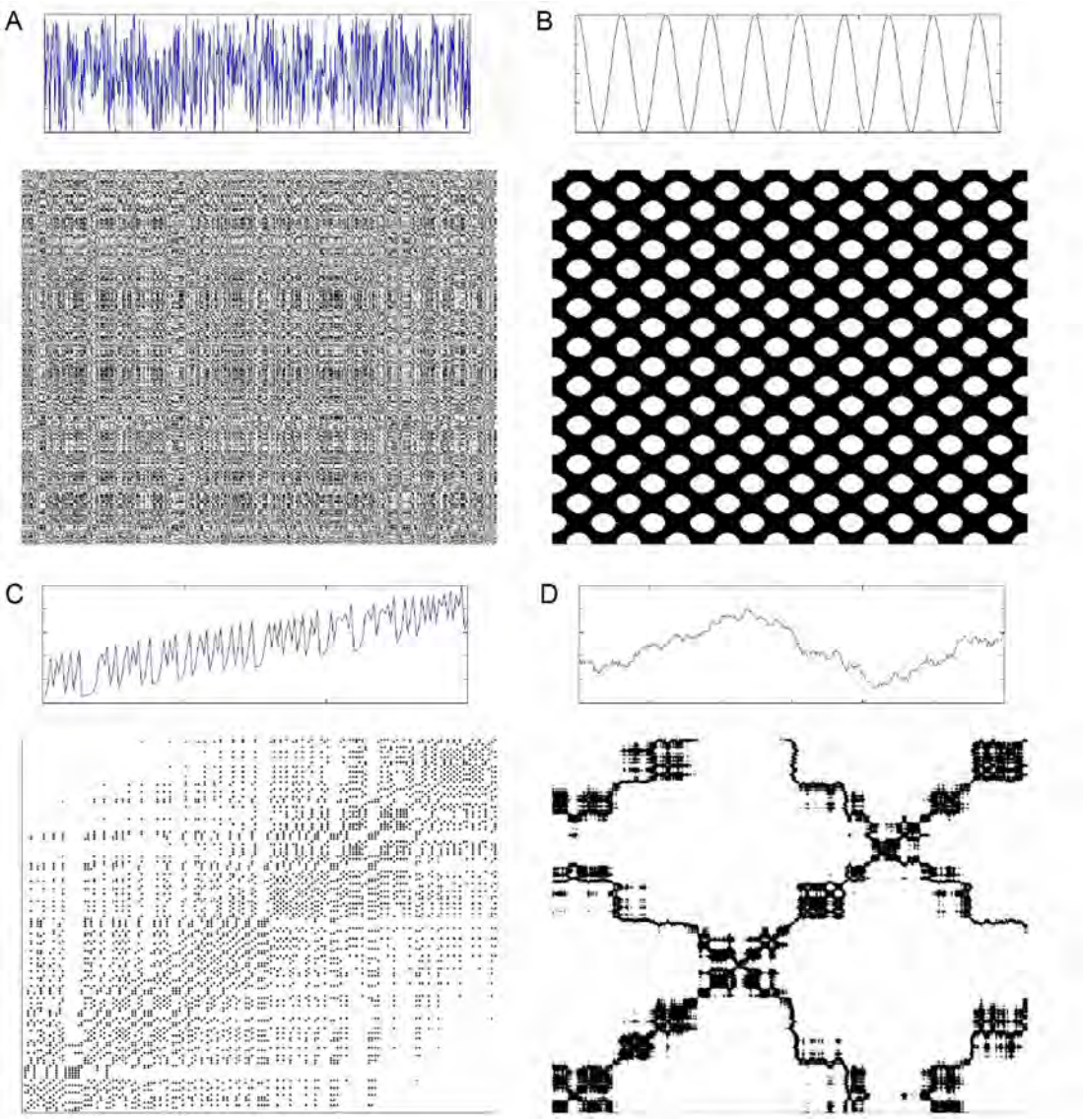

Figura 2.6 - Esses exemplos ilustram as tipologias características do gráfico de recorrência: (A) homogêneo: obtido a partir de uma série uniformemente distribuída com ruído branco; (B) periódico: obtido a partir da função $f(x) = sen(x)$; (C) obtido a partir do mapa logístico incrementado com um termo linear $x_{i+1} = 4x_i(1-x_i)+0,01i$; (D) descontínuo: obtido a partir do movimento browniano. Os dados usados possuem 600 pontos (A, B e D) e 150 (C). Os parâmetros usados para gerar os gráficos de recorrência são: $m = 1$ para todas as séries e $\epsilon = 0,2$ (A e C), $\epsilon = 1$ (B) e $\epsilon = 0,01$ (D), respectivamente. Adaptado de Marwan et al. (2007).

As estruturas de pequena escala (textura) podem ser classificadas em ponto isolado, linha diagonal, linhas vertical e horizontal (MARWAN et al., 2007):

- Pontos: isolados representam estados recorrentes e podem ocorrer em estados raros. Podem persistir por um tempo curto ou em fortes flutuações;
- Linha diagonal: ocorre quando um segmento da trajetória está em paralelo a outro segmento por um determinado tempo. Em outras palavras, representam trajetórias que evoluem do mesmo ϵ por um certo tempo;
- Linhas vertical e horizontal: marca um intervalo de tempo em que um estado não muda ou muda lentamente;

O gráfico de recorrência de uma série temporal é possível extrair informações sobre o sistema, mas essa interpretação visual requer alguma experiência. Essa interpretação visual é realizada a partir da compreensão da tipologia e textura presente nos gráficos de recorrência. Entretanto medidas de quantificação das estruturas obtidas são necessárias para uma investigação do sistema considerado (MARWAN et al., 2007).

Medidas de Quantificação

Além da interpretação visual dos gráficos de recorrência, são propostas diversas medidas de quantificação das pequenas estruturas (MARWAN et al., 2007; MARWAN et al., 2002) denominadas como *Análise de Quantificação de Recorrência*, em inglês *Recurrence Quantification Analysis - RQA*.

Essas medidas são baseadas na densidade das estruturas dos pontos recorrentes e de linhas diagonais e verticais nos gráficos de recorrência. Alguns estudos baseados em medidas RQA mostram que são capazes de identificar pontos de bifurcação, especialmente transição ordem-caos (TRULLA et al., 1996). As estruturas verticais no gráfico de recorrência podem estar relacionadas aos estados intermitentes e laminares e detectarem transições caos-caos (MARWAN et al., 2002).

Medida baseada na densidade recorrente

A medida RQA mais simples é a razão de recorrência (RR):

$$RR(\epsilon) = \frac{1}{N^2} \sum_{i,j=1}^{N} R_{i,j}(\epsilon) \tag{2.14}$$

que é a medida de densidade dos pontos recorrentes no gráfico de recorrência, onde N é o tamanho da série analisada. Quando $N \to \infty$, RR é a probabilidade do estado recorrer na vizinhança ϵ no espaço de estados.

Medidas baseadas nas linhas diagonais

Essas medidas são baseadas no histograma $P(\epsilon, l)$ de linhas diagonais de comprimento l:

$$P(\epsilon, l) = \sum_{i,j=1}^{N} (1 - R_{i-1,j-1}(\epsilon))(1 - R_{i+1,j+1}(\epsilon)) \prod_{k=0}^{l-1} R_{i+k,j+k}(\epsilon) \tag{2.15}$$

para simplificação das medidas RQA, o termo ϵ é omitido, usando $P(l) = P(\epsilon, l)$ (MARWAN et al., 2007).

As linhas diagonais são formadas quando o sistema percorre, em tempos distintos, uma mesma região do espaço de estados, de um mesmo modo. A existência de evoluções temporais similares é uma indicação da existência de regras determinísticas regendo o comportamento dinâmico do sistema. Por este motivo, a razão entre pontos de recorrências que formam estruturas diagonais e todos os pontos de recorrências é introduzida como uma medida para determinismo (DET) do sistema (LING, 2009):

$$DET = \frac{\sum_{l=l_{min}}^{N} lP(l)}{\sum_{l=1}^{N} lP(l)} \tag{2.16}$$

$P(l)$ é a probabilidade da estrutura diagonal ocorrer dentro do gráfico de recorrência, l_{min} é o tamanho mínimo de estruturas diagonais que se deseja contabilizar dentro do gráfico de Recorrência.

Uma linha diagonal de comprimento l significa que um segmento da trajetória durante l passos de tempo é próximo de uma outra trajetória em diferente tempo, logo, essas linhas são relacionadas à divergência dos segmentos das trajetórias. A média do comprimento da linha diagonal (L) que pode ser interpretado como o tempo médio de previsibilidade do sistema é dado por:

$$L = \frac{\sum_{l=l_{min}}^{N} lP(l)}{\sum_{l=l_{min}}^{N} P(l)} \tag{2.17}$$

Outra medida baseada nas linhas diagonais é a entropia ($ENTR$), referente à Entropia de Shannon da probabilidade $p(l) = P(l)/N_l$, onde $N_l = \sum_{l \geq l_{min}} P(l)$ é o número total de linhas diagonais, para encontrar uma linha diagonal de tamanho l no gráfico de Recorrência:

$$ENTR = -\sum_{l=l_{min}}^{N} p(l) \ln p(l) \tag{2.18}$$

Segundo Marwan et al. (2007) $ENTR$ reflete a complexidade do gráfico de recorrência em relação às linhas diagonais, por exemplo, um sinal não correlacionado possui um valor de $ENTR$ baixo, indicando baixa complexidade em comparação ao sinal correlacionado.

Medidas baseadas nas linhas verticais

O número total de linhas verticais, equivalente ao número total de linhas horizontais, já que o gráfico de recorrência é simétrico, de comprimento v é dado pelo histograma:

$$P(v) = \sum_{i,j=1}^{N} (1 - R_{i,j})(1 - R_{i,j+v}) \prod_{k=0}^{v-1} R_{i,j+k} \tag{2.19}$$

Semelhante ao determinismo (Equação 2.9), o raio entre os pontos recorrentes formam estruturas verticais e o conjunto dessas estruturas pode ser calculado pela medida denominada laminariedade (LAM):

$$LAM = \frac{\sum_{v=v_{min}}^{N} vP(v)}{\sum_{v=1}^{N} vP(v)} \tag{2.20}$$

Para mapas, $v_{min} = 2$ é um valor apropriado (MARWAN et al., 2007). A medida LAM decresce se o gráfico de recorrência consiste de mais pontos isolados recorrentes do que estruturas verticais. A média do comprimento das estruturas verticais e horizontais (TT) é dado por:

$$TT = \frac{\sum_{v=v_{min}}^{N} vP(v)}{\sum_{v=v_{min}}^{N} P(v)} \tag{2.21}$$

é denominado de tempo de aprisionamento (*trapping time*). TT mede o tempo médio que um estado permanece em um estado laminar, um estado que não muda no tempo (SOUZA, 2008).

2.3 Caracterização - técnicas de Mineração de Dados

Após calcular todos os valores dos índices dos conjuntos de séries temporais com os métodos de sistemas dinâmicos, esses valores servem como entrada para os classificadores. Esses classificadores exploram o conjunto de medidas (dados) dado, fornecendo informações e padrões que são associados às classes.

As técnicas de Mineração de Dados (MD) empregam, especialmente aquelas utilizadas nesse trabalho, um princípio de inferência denominado indução, que pode ser entendido como a capacidade de fazer conclusões genéricas a partir de um conjunto de exemplos (LORENA; CARVALHO, 2007). O aprendizado indutivo pode ser supervisionado ou não-supervisionado.

A diferença entre os dois tipos de aprendizagem é que no aprendizado supervisionado há um "professor", que apresenta seu conhecimento da forma da saída desejada, enquanto no aprendizado não supervisionado não há o conhecimento da saída desejada a partir de um determinado conjunto de exemplos.

O algoritmo supervisionado tem por meta fazer com que a representação gerada seja capaz de produzir as saídas corretas para entradas ainda não apresentadas. O algoritmo não supervisionado tem por objetivo aprender a representar as entradas segundo uma medida de qualidade, encontrando padrões ou tendências no conjunto de dados apresentados (LORENA; CARVALHO, 2007).

2.3.1 Classificadores de árvore de decisão

Classificadores de árvore de decisão possuem uma abordagem hierárquica. Classificadores hierárquicos são um tipo especial de classificadores multiestágio, ou seja, dividem uma decisão complexa em diversas decisões menores, esperando dessa maneira, que a solução final obtida seja semelhante à solução pretendida (SAFAVIAN; LANDGREBE, 1991). Esses classificadores podem ser usados em diferentes aplicações, como reconhecimento de caracteres, análise de sinais de radar, diagnóstico médicos, entre outras (ANDREOLA, 2009).

Algoritmos de árvore de decisão elaboram, a partir de um conjunto de dados, uma estrutura de dados em forma de árvore que pode ser usada para classificar novos

casos (SZALBERG, 1994) (ver Figura 2.7). Cada caso é descrito por um conjunto de características que podem ser valores simbólicos ou numéricos, associados ao seu rótulo que representa uma classe. Em outras palavras, é um fluxograma onde cada nó da árvore denota um teste de atribuição de valores, cada ramo representa uma saída para o teste e as folhas representam a classe ou distribuição de classes do conjunto de dados (HAN et al., 2011)

Para Safavian e Landgrebe (1991) uma árvore pode ser descrita como um grafo $G = (V, A)$ que consiste em um conjunto não vazio de vértices ou nós (V) interligado por um conjunto de arestas (A) da forma (v_i, v_j), sendo v_i e $v_j \in V$. Se as arestas são pares ordenados (v_i,v_j), do v_i para v_j o grafo é denominado direcionado. Uma árvore acíclica (sem ciclos) direcionada deve obedecer a algumas propriedades:

- possuir apenas um vértice ou nó raiz no nível 0, que contenha todos os padrões de todas as classes a serem classifcadas pelo classificador de decisão em árvore;
- todos os demais vértices, exceto a raiz, tem exatamente uma aresta de chegada havendo um único caminho da raiz aos demais nós;
- um vértice é denominado filho quando originado por determinado vértice que será chamado pai;
- os nós que não possuírem nós filhos são chamados de folha ou terminais. Nesses nós o padrão agora discriminado, recebe a identificação ou rótulo da classe em questão.

Há diversos algoritmos de árvore de decisão, sendo que o mais conhecido é o algoritmo ID3, desenvolvido por Quinlan (1986). A ideia conceitual do ID3 é apresentar uma janela (subconjunto de dados), escolhida aleatoriamente, do conjunto de treinamento e a partir dela elaborar uma árvore de decisão. Todos os elementos do conjunto são classificados usando essa árvore. Caso essa classificação seja realizada corretamente, o processo finaliza. Caso contrário, uma seleção dos objetos classificados incorretamente é adicionado à janela e o processo continua até que seja estabelecida uma árvore de decisão adequada para o conjunto de treinamento (ROCHA, 2008).

Considere uma coleção de C objetos da qual se pretende formar uma árvore. Caso essa coleção C seja vazia ou todos os objetos pertençam a mesma classe, a árvore será composta de apenas uma folha. Caso tenha-se mais de uma classe, sendo T um

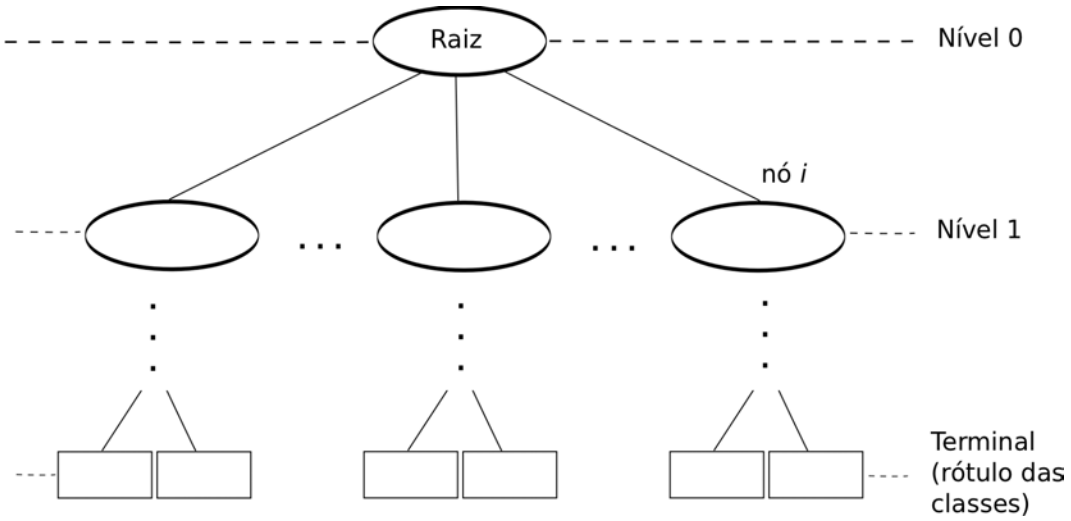

Figura 2.7 - Exemplo geral de árvore de decisão.
Fonte: Adaptado de Safavian e Landgrebe (1991)

teste qualquer sobre o objeto, os possíveis resultados poderão ser $O_1, O_2, ..., O_w$. Para cada objeto no conjunto C é obtido um T, produzindo uma partição $\{C_1, C_2, ..., C_w\}$ de C, com C_i com objetos de saída O_i.

A escolha do teste adequado pode ser baseada em duas hipóteses:

1*:* Toda árvore de decisão correta para C classificará objetos na mesma proporção que sua representação em C. Considere que uma amostra de objetos pertença a duas classes P e N, um objeto será da classe P com probabilidade $p/(p+n)$ ou classe N com probabilidade $n/(p+n)$, onde p é o total de objetos da classe P e n é o total de objetos da classe N.

2*:* Na classificação de um objeto, a árvore retorna uma classe. Para gerar a classificação P ou N utiliza-se:

$$I(p,n) = -\frac{p}{p+n}\log_2\left(\frac{p}{p+n}\right) - \frac{n}{p+n}\log_2\left(\frac{n}{p+n}\right) \tag{2.22}$$

Para que um atributo A com os valores $[A_1, A_2, ..., A_v]$ já considerado como o nó ou vértice raiz (nível 0) da árvore é necessário obter a média ponderada:

$$E(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p+n} I(p,n) \tag{2.23}$$

sendo p_i os objetos da classe P e n_i os objetos da classe N. O ganho de informação obtido por esse ramo usando o atributo A é dado por:

$$G(A) = I(p, n) - E(A) \tag{2.24}$$

O algoritmo escolhe o atributo A que maximiza o ganho de informação. O processo é recursivo formando a árvore para todos os subconjuntos de C, ou seja, $C_1, C_,2, ..., C_v$.

Para ilustrar a construção de uma árvore de decisão, considere o exemplo a seguir (Tabela 2.1), extraído de Witten e Frank (2005), sobre a decisão de jogar tênis ou não conforme a previsão do tempo:

Tabela 2.1 - Possibilidade de jogar tênis a partir dos dados sobre o tempo (WITTEN; FRANK, 2005).

Previsão	**Temperatura**	**Umidade**	**Vento**	**Jogar**
Ensolarado	Quente	Alta	Falso	Não
Ensolarado	Quente	Alta	Verdadeiro	Não
Nublado	Quente	Alta	Falso	Sim
Chuvoso	Média	Alta	Falso	Sim
Chuvoso	Fria	Normal	Falso	Sim
Chuvoso	Média	Normal	Verdadeiro	Não
Nublado	Fria	Normal	Verdadeiro	Sim
Ensolarado	Média	Alta	Falso	Não
Ensolarado	Fria	Normal	Falso	Sim
Chuvoso	Média	Normal	Falso	Sim
Ensolarado	Média	Normal	Verdadeiro	Sim
Nublado	Média	Alta	Verdadeiro	Sim
Nublado	Quente	Normal	Falso	Sim
Chuvoso	Média	Alta	Verdadeiro	Não

Considerando o conjunto composto por 14 amostras, 9 amostras são da classe SIM (classe P) e 5 são da classe NÃO (classe N). A informação requerida é:

$$I(9, 5) = -\frac{9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right) = 0,940\text{bits} \tag{2.25}$$

Para o atributo *previsão* há os rótulos: *ensolarado* (2 para classe P e 3 para classe N), *chuvoso* (4 para classe P e 0 para classe N) e *nublado* (3 para classe P e 2 para classe N). Calculando a informação para cada rótulo do atributo *previsão*, sendo

$I(p_1, n_1)$ o rótulo *ensolarado*, $I(p_2, n_2)$ o rótulo *nublado* e $I(p_3, n_3)$ o rótulo *chuvoso*:

$$I(p_1, n_1) = \frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right) = 0,971\text{bits} \quad (2.26)$$

$$I(p_2, n_2) = 0\text{bits} \quad (2.27)$$

$$I(p_3, n_3) = 0,971\text{bits} \quad (2.28)$$

A informação necessária depois da verificação desse atributo é $E(\text{previsão}) = 0.694\text{bits}$ e o ganho de informação é $G(\text{previsão}) = 0.246\text{bits}$. Assim, o ganho de informação para os demais atributos é: $G(\text{temperatura}) = 0.029\text{bits}$, $G(\text{umidade}) = 0.151\text{bits}$ e $G(\text{vento}) = 0.048\text{bits}$.

Analisando os ganhos obtidos para os atributos, o atributo *previsão* obtém o maior ganho, logo é escolhido como nó raiz da árvore. O processo é recursivo para os demais nós da árvore. A árvore final obtida dessa classificação é apresentada na Figura 2.8.

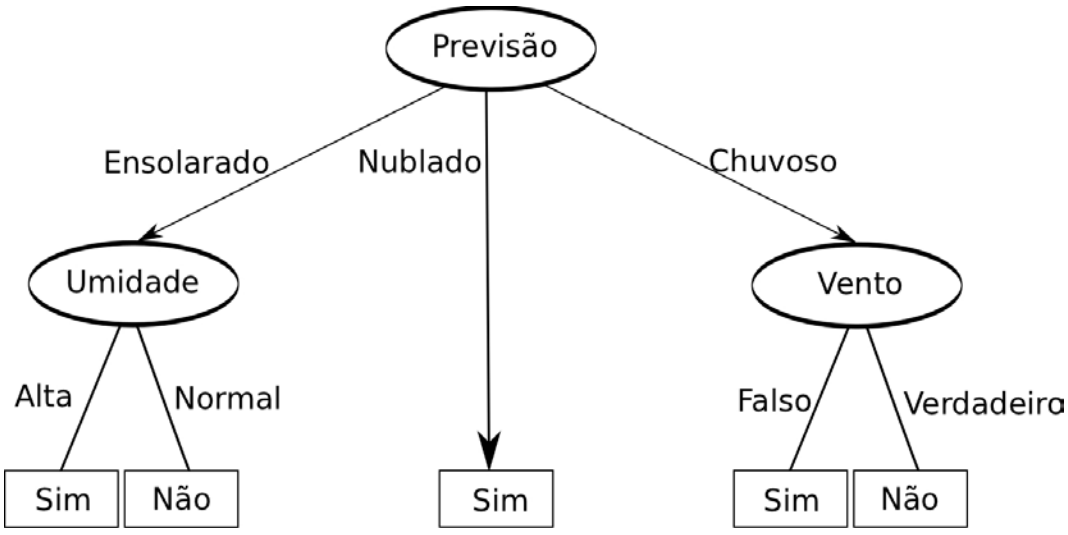

Figura 2.8 - Exemplo geral de árvore de decisão.
Fonte: Adaptado de Witten e Frank (2005)

2.3.2 Máquinas de Vetores de Suporte

Máquinas de Vetores de Suporte ou *Support Vector Machines* (SVM) são um método de aprendizado de máquina usado para classificação, regressão ou outras tarefas de

aprendizado (NATIONAL TAIWAN UNIVERSITY, 2012). A técnica SVM foi desenvolvida por Cortes e Vapnik (1995).

Os resultados desse método de classificação são, muitas vezes, superiores aos resultados obtidos com outras técnicas de classificação de dados por aprendizado, como as Redes Neurais Artificiais (RNAs) (LORENA; CARVALHO, 2007). Esse método pode ser aplicado na solução de problemas relacionados à categorização de textos, análise de imagens e bioinformática (WANG, 2002).

A principal ideia dessa técnica é utilizar um hiperplano linear de separação que maximiza a distância entre duas classes para um classificador, denominado margem. Se o problema não for linearmente separável, o método SVM emprega duas abordagens para solução. A primeira abordagem é construir um hiperplano de margens suaves que adiciona uma função de penalidade de violação no critério de otimização.

A segunda abordagem é transformar não linearmente o espaço de entradas original para uma dimensão maior no espaço de características. Então, nesse novo espaço de características é mais provável encontrar um hiperplano linear de separação ótimo (ver Figura 2.9) (WANG, 2002; ARAUJO, 2010).

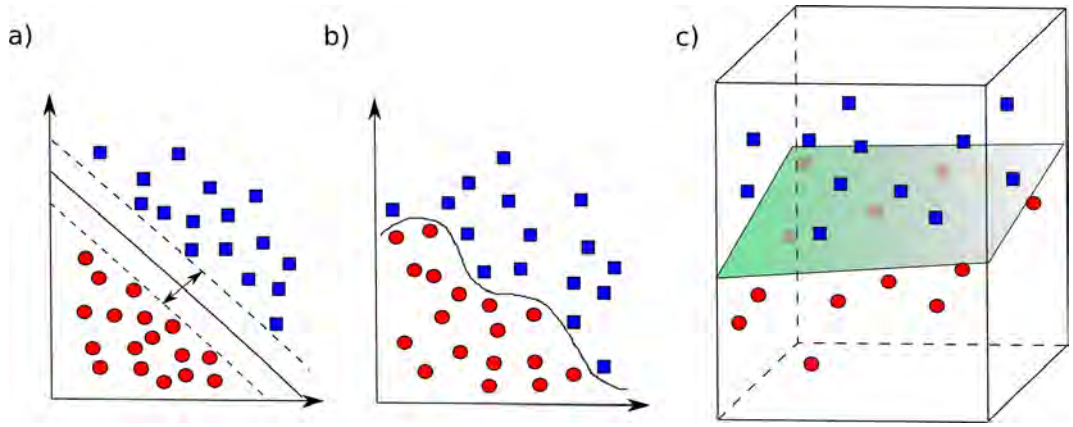

Figura 2.9 - Exemplo de separação de duas classes de um conjunto pelo SVM. a) Formação da margem quando duas amostras são linearmente separáveis, b) duas classes que não são linearmente separáveis e que considerando uma dimensão mais elevada (c) permite a construção de um hiperplano para separação das classes.

Considere os vetores de treinamento $x_i =$, $i = 1, ..., l$ de comprimento n, um vetor y

é definido como:

$$y_i = \begin{cases} 1 \text{ se } x_i \text{ para classe } 1, \\ -1 \text{ se } x_i \text{ para classe } 2 \end{cases} \tag{2.29}$$

A técnica SVM tenta encontrar o hiperplano separando com maior margem as duas classes, por exemplo na Figura 2.9a as duas classes podem ser totalmente separadas por uma linha $w^T x + b = 0$. Para definir a maior margem entre as duas classes é preciso estabelecer uma linha com parâmetro w e b tal que a distância $w^T x + b = \pm 1$ é maximizada. Como a distância entre $w^T x + b = \pm 1$ é $2/||w||$ e a maximização $2/||w||$ é equivalente $w^T w/2$, há o seguinte problema (WANG, 2002):

$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2} w^T w \\ & y_i((w^T x_i) + b) \geq 1 \\ & i = 1, ..., l. \end{aligned} \tag{2.30}$$

A restrição $y_i((w^T x_i) + b) \geq 1$ pode ser definida como:

$$\begin{aligned} (w^T x_i) + b \geq 1 \text{ se } y_i = 1, \\ (w^T x_i) + b \leq -1 \text{ se } y_i = -1. \end{aligned} \tag{2.31}$$

Os dados da classe 1 ficam do lado direito da $w^T x + b = 0$ e os da outra classe do lado esquerdo. A razão de maximização da distância entre $w^T x + b = \pm 1$ é baseada na minimização do risco estrutural de Vapnik (VAPNIK, 1998).

Na prática, a maioria dos problemas são não linearmente separáveis (Figura 2.9b). Para resolver esse tipo de problema, SVM usa dois métodos: primeiramente as SVM permitem erros de treinamento. E em segundo, há transformação não linear do espaço de entradas originais em um espaço de características de dimensão maior dada uma função ϕ:

$$\min_{w,b,\xi} \quad \frac{1}{2} w^T w + C\Big(\sum_{i=1}^{l} \xi_i\Big) \tag{2.32}$$

$$y_i((w^T \phi(x_i)) + b) \geq 1 - \xi_i, \tag{2.33}$$

$$\xi_i \geq 0, \ i = 1, ..., l.$$

Um termo de penalidade C $\left(\sum_{i=1}^{l} \xi_i\right)$ na função objetivo e erros de treinamento são permitidos. Ou seja, a restrição (Equação 2.33) permite que os dados de treinamento possam não estar do lado correto do hiperplano, enquanto minimiza o erro de treinamento $\sum_{i=1}^{l} \xi_i$ na função objetivo (WANG, 2002). Logo, se o parâmetro C é suficientemente grande e o dado é linearmente separável, o problema de restrição da Equação 2.33 volta para a restrição da Equação 2.30 e ξ_i será zero. Note que os dados de treinamento x é mapeado em um vetor (possivelmente infinito) de espaço de maior dimensão:

$$\phi(x) = (\phi_1(x), \phi_2(x), ...) \tag{2.34}$$

A Equação 2.32 é um problema de espaço dimensional infinito que não é fácil de resolver. O processo usual de solução é resolvendo a dupla formulação da Equação 2.32, que precisa de aproximar da forma $k(x_i, x_j) \equiv \phi(x_i)^T \phi(x_j)$ que é denominado função *kernel*. Alguns tipos de kernel conhecido são, por exemplo, o kernel RBF ($e_{-\gamma||x_i - x_j||^2}$) e o kernel polinomial ($(x_i^T x_j/\gamma + \delta)^d$), onde γ e δ são parâmetros (WANG, 2002).

Depois de solucionado a forma dual, a função de decisão é reescrita:

$$f(x) = \text{sign}(w^T \phi(x) + b) \tag{2.35}$$

Em outras palavras, para o vetor teste x se $w^T \phi(x) + b > 0$, classificamos como classe 1, caso contrário como classe 2. Somente alguns x_i, $i = 1, ..., l$ são usados para construir w e b e são denominados de *vetores de suporte*. Portanto, dizemos que SVM é usado para encontrar dados importantes, os vetores suporte, para o treinamento dos dados.

Utilização das SVM

Nesse trabalho, o pacote LIBSVM (NATIONAL TAIWAN UNIVERSITY, 2012) é usado. Está disponível em `http://www.csie.ntu.edu.tw/~cjlin/libsvm`. Esse pacote pode ser usado para classificação binária de classes e de multi-classes, diferentemente da versão tradicional que permite a classificação de duas classes. Entretanto, nesse estudo usamos comparações de duas classes distintas.

Para a análise do conjunto de medidas fornecidas pelas técnicas de sistemas dinâmicos usando o classificador SVM são adotadas três abordagens diferentes relacionadas aos padrões de entrada. A primeira abordagem é a apresentação de todas as medi-

das de cada comparação entre grupos obtendo um resultado. O resultado do grupo de teste fornecido pelo classificador, ou seja, a capacidade de acertar a qual grupo pertencia determinado conjunto de medidas é denominado de acurácia. A acurácia foi definida como a quantia de acertos no grupo de teste dividido pelo número total de casos de teste obtendo-se a acurácia média no final de todas as execuções para cada comparação entre dois grupos (realiza-se 100 execuções).

A segunda abordagem é apresentar apenas uma medida por vez ao classificador que após ser executado 100 vezes calcula-se a acurácia média. Essa segunda abordagem permite estabelecer quais medidas são capazes ou não de diferenciar entre os dois grupos comparados, detectando dinâmicas diferentes.

A terceira abordagem é apresentar apenas duas medidas por vez ao classificador, estimando a acurácia média que é associada às duas medidas fornecidas como entrada. Foram testadas todas as combinações dois a dois de medidas para as comparações dos grupos realizadas.

Para ilustrar essa terceira abordagem, a Figura 2.10 apresenta essa estratégia que compreende três etapas. Na Etapa 1 uma combinação de duas medidas (A e B) são fornecidas como a entrada para o classificador. Nessa combinação das medidas constam casos pertencentes às duas classes ou grupos de pacientes (casos 1 a n para as classes 1 e 2), estabelecidos aleatoriamente.

Sendo assim o grupo de treinamento é formado por: $A_{1,1}$,...,$A_{1,n}$ e $A_{2,1}$,...,$A_{2,n}$ para os valores da medida A das classes 1 e 2 e $B_{1,1}$,...,$B_{1,n}$ e $B_{2,1}$,...,$B_{2,n}$ para os valores da medida B das classes 1 e 2. O processo de aprendizagem é supervisionado e a saída durante o treinamento do classificador é uma resposta esperada de acordo com os rótulos das classes fornecidos, ou grupo 1 ou grupo 2.

Na Etapa 2, após treinamento, um novo conjunto de entrada, o de teste, é fornecido ao classificador. Esse grupo de teste, também escolhido aleatoriamente, é composto pelos mesmos tipos de medidas (A e B) e pelo mesmos grupos de pacientes, entretanto, outros casos que não são usados durante o treinamento. A partir da classificação fornecida pelo algoritmo, calcula-se a acurácia média (o classificador é executado 100 vezes) que é associada as duas medidas de entrada. No exemplo, o valor de acurácia média é $0,6$.

A Etapa 3 ilustra a elaboração do gráfico para visualização dos valores obtidos em cada medida, considerando todas as combinações dois a dois. No exemplo da

Figura 2.10, o valor de acurácia $0,6$, pertence tanto para a medida A como para B. Considerando outros exemplos para combinações dois a dois de medidas (B e C; B e D; C e D; A e C; A e D) temos: as acurácias $0,3$ para as medidas B e C; $0,32$ para as medidas B e D; $0,6$ para C e D; $0,9$ para A e C e 1 para A e D.

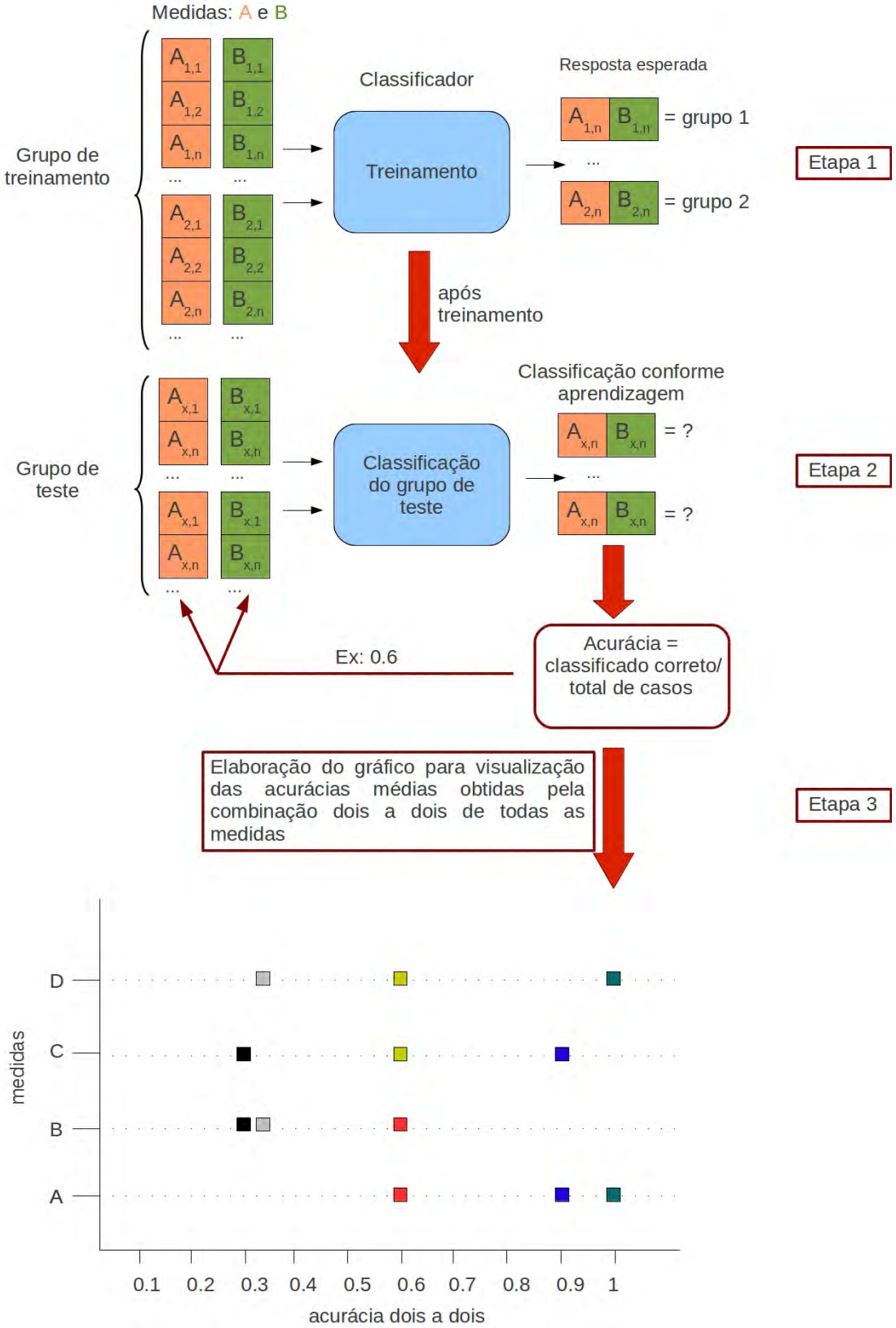

Figura 2.10 - Ilustração da terceira abordagem para caracterização das séries de intervalos RR usando o SVM. Etapa 1 há o treinamento do classificador, na Etapa 2 a classificação do grupo de teste e na Etapa 3 cálculo da acurácia média associada as duas medidas de entrada e a elaboração do gráfico do desempenho de todas as medidas.

3 SÉRIES TEMPORAIS DE INTERVALOS RR

Nesse Capítulo é descrito o primeiro estudo de caso usado na metodologia proposta apresentada no Capítulo 2. Aqui estão descritas as séries temporais de intervalos RR ou tacogramas que são oriundas do ritmo cardíaco, a interrelação entre o coração (ritmo cardíaco) e o sistema nervoso autônomo, algumas limitações do uso desse tipo de série temporal, os bancos de dados utilizados e o tipo de pré-processamento específico neste conjunto de sinais.

3.1 Dinâmica de variação do batimento cardíaco

A frequência cardíaca ou ritmo cardíaco pode ser definida como o número de vezes que o coração bate[1] em um determinado intervalo de tempo. Entretanto essa frequência no número de batimentos cardíacos pode variar conforme cada indivíduo e o seu ambiente, faixa etária ou ainda condições físicas e de saúde.

Alguns autores sugerem que essa irregularidade no ritmo cardíaco seria devido a sua possível dinâmica caótica. Entretanto, identificar a presença de caos em sistemas biológicos (séries temporais experimentais) é desafiador pois requer uma prova definitiva da presença de determinismo nas séries analisadas e ainda não é possível tal prova. É possível detectar algumas propriedades das séries permitindo obter informações relevantes (FREITAS et al., 2009). Em relação à análise de séries temporais obtidas de sistemas cardíacos, muitos trabalhos usam argumentos geométricos como dimensão de correlação e expoentes de Lyapunov para identificar caos em freqüência cardíaca, porém, nenhum desses trabalhos é muito conclusivo em relatar a presença ou não de caos nos ritmos cardíacos.

Segundo Freitas et al. (2009) antes de afirmar que um comportamento é caótico, deve haver uma clara evidência de que as equações deterministas governam a sua dinâmica. Um dos primeiros artigos dedicados a identificar determinismo na freqüência cardíaca foi Pool (1989) e que essa ainda é considerada uma questão aberta (existência de determinismo na freqüência cardíaca) (REDDY; KUNTAMALLA, 2011).

Uma forma de estudar a dinâmica da frequência cardíaca é como esta varia ao longo do tempo. Analisar a variabilidade da freqüência cardíaca (VFC) é uma medida simples e não invasiva dos impulsos autônomos (relacionados ao sistema nervoso autônomo). Essa variabilidade é normal e esperada, indicando a habilidade do coração em responder aos múltiplos estímulos fisiológicos e ambientais (respiração,

[1] O batimento do coração é uma das funções do sistema nervoso autônomo.

exercício físico, estresse mental, alterações hemodinâmicas e metabólicas, sono e desordens induzidas por doenças) (KUUSELA, 2013).

A VFC descreve as oscilações dos intervalos entre batimentos cardíacos consecutivos (intervalos RR) permitindo analisar as influências do sistema nervoso autônomo (SNA). Mudanças nos padrões da VFC podem indicar a sensibilidade do SNA e possíveis comprometimentos de saúde, uma alta VFC indica uma boa adaptação do indivíduo saudável com mecanismos autonômicos eficientes e uma baixa VFC mostra uma possível adaptação anormal e insuficiente do SNA (podendo ser presença de mau funcionamento fisiológico no indivíduo) (VANDERLEI et al., 2009).

A forma de se obter a VFC é através do *eletrocardiograma* (ECG). O ECG é um método não invasivo e de fácil utilização. Os instrumentos de captação do ECG são denominados de eletrocardiógrafos (conversores analógicos ou digitais) e os cardiofrequencímetros (sensores externos colocados em pontos específicos do corpo) (VANDERLEI et al., 2009). Para analisar as ondas obtidas pelo ECG é necessário compreender primeiramente como ocorre a relação entre o coração e o sistema nervoso autônomo.

3.2 Relação entre o coração e sistema nervoso autônomo

Regular o funcionamento do coração é uma das funções do sistema nervoso autônomo (SNA)[2]. O SNA é responsável por regular os processos fisiológicos do organismo humano, estabelecendo, portanto, o que se denomina de condições normais e condições patológicas do organismo (VANDERLEI et al., 2009). Entre esses processos fisiológicos podemos destacar a respiração, circulação do sangue, controle da temperatura e da digestão e controle dos ritmos cardíacos.

O SNA está interligado ao hipotálamo[3], que dentre suas características coordena a resposta comportamental do organismo para garantir a homeostasia[4]. O SNA pode ser dividido em dois sistemas: o sistema nervoso simpático (SNS) e sistema nervoso parassimpático (SNP). Essa divisão é baseada nas características anatômicas de cada parte e nas funções fisiológicas que cada uma delas desempenha.

[2]O sistema nervoso autônomo é constituído por um conjunto de neurônios que se encontram na medula e tronco encefálico. Esse, através de gânglios periféricos, coordena diversas atividades no organismo.

[3]Hipotálamo é uma região do encéfalo que é responsável entre diversas características por regular processos metabólicos, atividades autônomas; pela secreção de neuro hormônios; pelo controle da temperatura corporal, fome, sede e ciclos circadianos (que levam 24 horas) etc.

[4]Homeostasia é a propriedade de um sistema aberto (seres vivos) de regular o seu ambiente interno para manter uma condição estável.

O SNS localiza-se na região toracolombar (ver Figura 3.1), onde os nervos se localizam na região torácica e lombar medular, estimula ações que permitem ao organismo responder a situação de estresse. Dentre as diversas funções, é responsável pela aceleração dos batimentos cardíacos, aumento da pressão arterial, incremento da concentração de açúcar no sangue e ativação do metabolismo geral do corpo. O SNP localizado na região craniossacral (ver Figura 3.1) onde os nervos se localizam no tronco cerebral e na medula sacral, permite ao organismo responder a situação de calma. Dentre as diversas funções, é responsável pela desaceleração dos batimentos cardíacos, decréscimo da pressão arterial, diminuição da adrenalina e redução do nível de açúcar no sangue.

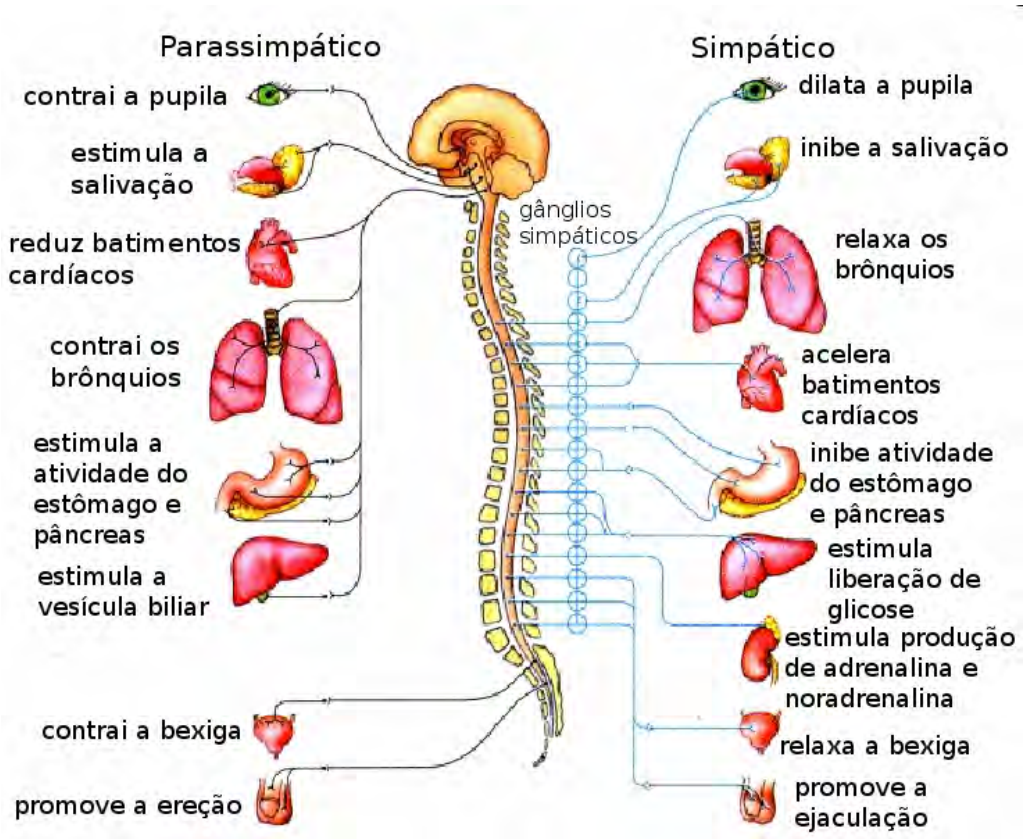

Figura 3.1 - Representação do sistema nervoso autônomo e suas divisões.

Como ser visto na Figura 3.1 em resposta às variações do ambiente o SNA para as variações do ambiente modifica os batimentos cardíacos (acelerando-o ou desacelerando-o) fazendo, por conseguinte, com que a VFC também se altere.

Para promover os batimentos cardíacos, o *coração*[5] é provido de um conjunto de ramificações nervosas, que constituem o tecido de condução do impulso nervoso. Esse tecido de condução é formado pelos nódulos sinoatrial (localizado no átrio direito do coração), atrioventricular (localizado entre o átrio e o ventrículo) e o feixe de His (ramificação espalhada por algumas partes do coração - que conduz o impulso nervoso para que ocorra a contração cardíaca) (ver Figura 3.2).

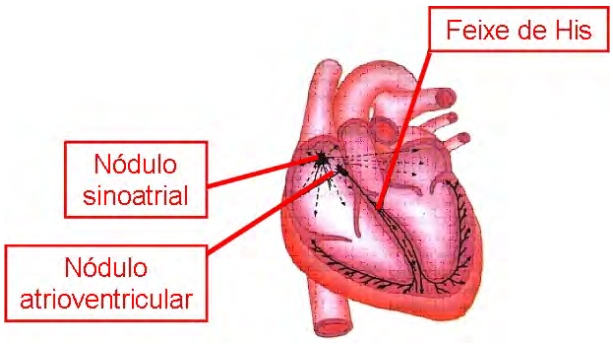

Figura 3.2 - Representação do coração evidenciando o tecido de condução nervosa.

Com o ECG é possível detectar a atividade do coração através dos batimentos cardíacos. A Figura 3.3 mostra as ondas capturadas pelos sensores e que representam a dinâmica cardíaca.

A excitação cardíaca inicia-se pelo impulso gerado no nódulo sinoatrial, distribuído pelos átrios denominado de *despolarização*[6] atrial ou simplesmente excitação atrial. No ECG essa excitação é representada pela onda P. Esse impulso é conduzido aos ventrículos pelo nódulo atrioventricular e distribuído pelas fibras de His (despolarização ventricular).

Essa despolarização ventricular é representada no ECG pelas ondas Q, R e S denominado complexo QRS. Com a condução do impulso até os ventrículos, para a despolarização ventricular, ocorre a sístole, ou seja, esvaziamento dos átrios enviando sangue para os ventrículos e diástole atrial, ou seja, intumescimento dos átrios.

[5]O coração é o órgão central da circulação, localizado na caixa torácica, constituído por duas porções: porção direita por onde circula sangue venoso (rico em gás carbônico) e a porção esquerda por onde circula sangue arterial (rico em oxigênio). Cada porção é formada por duas partes, a superior é o átrio (que recebe sangue de todas as partes do corpo ou dos pulmões) e a inferior é o ventrículo (que envia sangue para todas as partes do corpo ou para os pulmões).

[6]A despolarização de uma célula se refere à saída de repouso pela entrada de íons de $Na+$ na célula, sendo assim, em uma célula cardíaca, corresponde à contração muscular.

A *repolarização*[7] ventricular é representado no ECG pela onda T. Os intervalos entre as ondas R e R são denominados intervalos RR ou batimentos cardíacos consecutivos que correspondem à frequência de despolarização ventricular, cuja variação é denominada VFC.

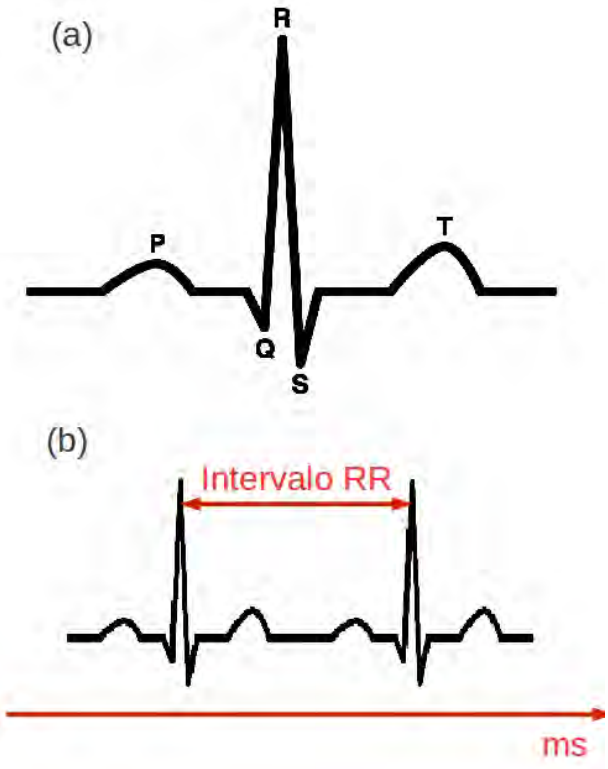

Figura 3.3 - a) Representação das ondas obtidas em um eletrocardiograma (ECG). b) Representação do intervalo RR entre duas ondas R, que corresponde a um batimento cardíaco completo.

Portanto, analisar a VFC contribui na avaliação do equilíbrio entre as influências simpáticas e parassimpáticas no ritmo cardíaco. A VFC pode ser usada para compreensão de diversas condições, dentre elas podem ser citadas: doença arterial coro-

[7] A repolarização de uma célula se refere ao período em que ela está voltando a sua polaridade normal (potencial de repouso), sendo assim, em uma célula cardíaca, corresponde ao relaxamento muscular.

nariana[8] (CARNEY et al., 2007), cardiomiopatia hipertrófica[9] (BITTENCOURT et al., 2010).

É importante salientar que essa análise é limitada a algumas situações fisiológicas ligadas ao organismo analisado. A seguir, algumas dessas situações são abordadas.

3.2.1 Limitações da análise da variabilidade da freqüência cardíaca

A análise da VFC pode contribuir na avaliação do equilíbrio entre as influências simpáticas e parassimpáticas do ritmo cardíaco desde que certas condições sejam respeitadas. Sabe-se que diante de algumas situações a VFC não pode ser associada às funções simpáticas e parassimpática do SNA. Segundo (VANDERLEI et al., 2009) algumas situações em que essa relação não pode ser considerada são:

- em pacientes transplantados, a análise da VFC não representa a modulação do coração pelo SNA, pois, o controle do coração denervado é feito em função do retorno venoso, da estimulação de receptores atriais, do estiramento atrial e de hormônios e outras substâncias presentes no sistema circulatório;
- em pacientes que possuem marcapasso cardíaco artificial, a análise da VFC não representa a modulação autonômica do coração, pois, o marcapasso é um dispositivo de estimulação multiprogramável capaz de substituir impulsos elétricos e/ou ritmos ectópicos, para se obter atividade elétrica cardíaca a mais fisiológica possível (ou seja, os pacientes possuem sua VFC modulada por esse dispositivo eletrônico);
- em pacientes com bloqueio atrioventricular, a análise da VFC não representa de forma adequada os intervalos RR uma vez que, esse bloqueio se caracteriza como um distúrbio de condução elétrica através do nódulo atrioventricular.

Pode-se analisar a VFC em diversas condições ambientais. Uma condição de extrema relevância e que está ligada ao contexto espacial é a mudança na VFC causada pela microgravidade (MARTINELLI et al., 2009). Sabe-se que a microgravidade

[8]Doença arterial coronariana também denominada de arterosclerose coronariana é caracterizada pelo estreitamento dos vasos que suprem o coração em decorrência do espessamento da camada interna da artéria devido ao acúmulo de placas.

[9]Cardiomiopatia hipertrófica é uma doença do miocárdio (músculo do coração) na qual uma porção do miocárdio está hipertrofiada (espessada) sem nenhuma causa óbvia. É a causa mais comum de morte súbita em atletas jovens.

causa mudanças na resposta adaptativa humana, principalmente com respeito ao sistema cardíaco. Entretanto, segundo Guerra (2008) a análise das séries cardíacas em condições de microgravidade ainda carece de abordagens mais sofisticadas que envolvam tanto técnicas avançadas de análise de sinais, como também a elaboração de modelos teóricos que simulem a geração de sinais relacionados à dinâmica cardíaca em seus mais variados estados.

Uma limitação da análise da VFC em séries temporais experimentais oriundas em microgravidade está ligada ao fato de que os experimentos de microgravidade em sua maioria são realizados em terra (como os vôos parabólicos) e só conseguem fornecer poucos segundos $(20 - 100s)$ em gravidade reduzida ou zero. Isso faz com que algumas técnicas de análise possam ser comprometidas devido ao tamanho das séries temporais experimentais.

3.3 Bancos de dados

O conjunto de séries temporais de intervalos RR ou tacogramas para realização desse estudo de caso foram obtidas de indivíduos em diferentes situações clínicas. Optou-se por uma casuística diversificada do ponto de vista clínico justamente para poder-se avaliar a aplicabilidade do método proposto diante da vasta gama de condições clínicas existentes em uma situação de mundo real. Foi possível assim, dispor de um amplo leque de tacogramas originários de pacientes, desde recém-nascidos prematuros até pacientes adultos idosos coronariopatas cirúrgicos, passando pelos adultos jovens sadios.

Conjuntos de séries temporais de intervalos RR de três bancos de dados distintos são usados nesse estudo de caso:

- provenientes do Núcleo Transdisciplinar de Estudos de Complexidade e Caos (NUTECC) sediado na Faculdade de Medicina de São José do Rio Preto, na cidade de São José do Rio Preto, São Paulo;

- provenientes do *Complexe de Recherche Interprofessionnel en Aerothermochimie* (CORIA) ligado à *Université de Rouen*, na cidade de Rouen, Normandia, França;

- provenientes do Banco de Dados PhysioNet, disponível gratuitamente na internet.

3.3.1 NUTECC

O conjunto de dados do NUTECC é composto das seguintes séries temporais de intervalos RR:

- 53 recém-nascidos prematuros (RNP) internados em unidades de terapia intensiva (UTI Neonatal) visando estabilização até a aquisição de condição de alta hospitalar com média de idade corrigida pelo método de Parkins de menos $27,4$ dias (SELIG et al., 2011);
- 26 recém-nascidos normais (RNN) (SELIG et al., 2011);
- 61 adultos jovens sadios (VOL) com média de idade de $20,7 \pm 1,6$ anos, em boas condições de saúde e praticantes de atividade física regular, oriundos de um grupo de estudo sobre reabilitação física de um centro universitário;
- 41 adultos sob dieta de muito baixo valor calórico para perda de peso (ADC), com média de idade de $48,9 \pm 16$ anos, que procuraram uma clínica de reabilitação alimentar visando melhora de sua condição nutricional, geralmente para repouso e/ou emagrecimento ;
- 61 adultos em avaliação pré-operatória para cirurgia de revascularização cirúrgica do miocárdio por coronariopatia obstrutiva grave (PC), com média de idade de $58,4 \pm 10,2$ anos (GODOY et al., 2005);
- 88 crianças de 8 a 13 anos com peso normal considerando seu índice de massa corpórea[10] (CONT) (VANDERLEI et al., 2010);
- 88 crianças de 8 a 13 anos com sobrepeso considerando seu índice de massa corpórea (COB) (VANDERLEI et al., 2010);

Foram obtidas seqüências de duração variável (em torno de 15 minutos até cerca de 1 hora), com o paciente em decúbito dorsal horizontal, procurando-se evitar estimulação externa sonora ou visual. Os equipamentos utilizados para captação foram os cinto Polar (S810i ou RS 800) já validados como adequados para captação de batimentos cardíacos no estudo da variabilidade da freqüência cardíaca (GAMELIN et al., 2006; VANDERLEI et al., 2008; NUNAN et al., 2009).

[10] Índice de massa corpórea (IMC) é um índice que estabelece uma relação entre o peso corpóreo e a altura do indivíduo. É considerado os seguintes valores: $< 18,5$ abaixo do peso, $18,6 - 24,9$ saudável, $25 - 29,9$ sobrepeso, acima de 30 obesidade.

Esse sistema comercial é adequado para medida de índices da variabilidade da frequência cardíaca em atendimento às recomendações de acordo com Task Force of the European Society of Cardiology the North American Society of Pacing Electrophysiology (1996). Esse dispositivo captura os intervalos RR em uma taxa de amostragem de 1000 Hz, por meio de eletrodos ligados a uma tira elástica colocada em torno do tórax (ver Figura 3.4). Todos estudos realizados envolvendo esse conjunto de dados obtiveram aprovação pelos seus respectivos comitês de ética.

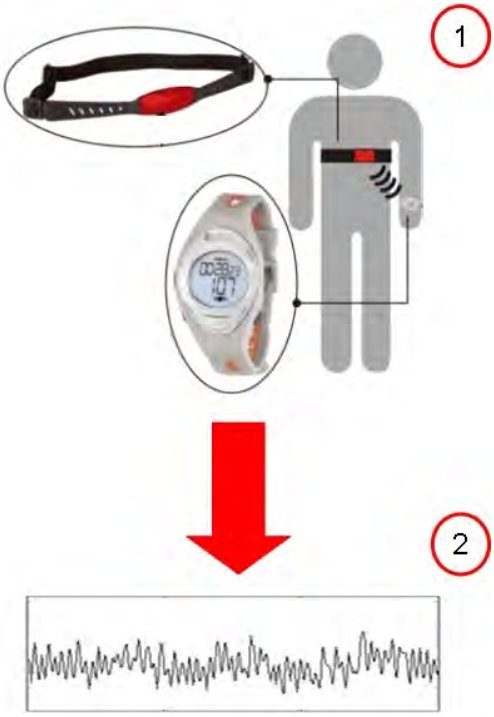

Figura 3.4 - Representação em duas etapas da captação da frequência cardíaca de um indivíduo usando o cinto Polar: 1 indivíduo usando o cinto e o relógio Polar e 2 série temporal de intervalos RR fornecida pelo equipamento.

3.3.2 CORIA

Os dados utilizados do banco de dados do CORIA são provenientes de pacientes submetidos a ventilação não invasiva noturna devido à insuficiência respiratória crônica (APN). Ventilação não invasiva é uma forma de tratamento para diagnósticos de insuficiência respiratória crônica que, por pressão positiva, corrige o mecanismo ventilatório (NAECK, 2011).

Esse conjunto de dados é constituído de séries temporais de RR obtidos de ECG (detecção dos intervalos RR por algoritmo disponível em `www.physionet.org`) em

polissonografia. Captado em momentos distintos, ao todo são nove pacientes que realizaram polissonografias antes de utilizar a ventilação, na primeira e na décima quinta noite após iniciar o procedimento, totalizando 27 séries temporais de intervalos RR.

3.3.3 PhysioNet

O PhysioNet (*www.physionet.org*) (GOLDBERGER et al., 2000) é um banco de dados de domínio público que disponibiliza além de diversas ferramentas para análise de sinais biológicos, diversos tipos de sinais biológicos. Entre os sinais disponibilizados estão séries de intervalos RR.

Os sinais usados nesse estudo de caso foram obtidos pelo equipamento Holter que monitora o indivíduo ao longo de 24 horas, fornecendo as séries de intervalos RR a partir do eletrocardiograma. São usadas para esse estudo de caso:

- 54 séries temporais de intervalos RR de adultos com ritmo sinusal normal (NOR) (BIGGER et al., 1995; STEIN et al., 1999);
- 15 séries temporais de intervalos de adultos com falha congestiva cardíaca severa (CHF) (BAIM et al., 1986).

3.4 Pré-processamento das séries

Para este primeiro estudo de caso, análise das séries temporais de intervalos RR, o tipo de pré processamento realizado é a filtragem dessas séries. O objetivo da filtragem é a remoção dos artefatos presentes nas séries, que poderiam conduzir erroneamente a interpretação dos resultados obtidos, uma vez que, conforme já mencionado, estamos interessados na análise do ritmo sinusal.

Há na literatura muitos trabalhos sobre filtragem da frequência cardíaca e outros métodos de pré processamento, análise de frequência cardíaca prenatal (fetal) (MANTEL et al., 1990; BERNARDES et al., 1991; GONCALVES et al., 2006b) e isso é justificado porque a análise de VFC é massivamente e universalmente usada durante a vida fetal, depois do nascimento (GONCALVES et al., 2007; Ayres-de-Campos et al., 2005) e durante o trabalho de parto (GONCALVES et al., 2006a; COSTA et al., 2009). Aqui, analisamos as séries temporais de intervalos RR (conforme já definido anteriormente) pós natal de indivíduos, incluindo recém nascidos a adultos.

São descritos somente dois tipos de filtragem nesse estudo: a convencional e a adap-

tativa. A filtragem convencional é realizada somente nos dados do NUTECC pelo especialista, com finalidade de comparar as séries obtidas dessa filtragem com a proposta para automatizar o procedimento de filtragem, que é a adaptativa.

3.4.1 Filtragem convencional

O processo de filtragem convencional é aplicado ao banco de dados do NUTECC e para os demais bancos é usado apenas o método de filtragem adaptativa. O método de filtragem convencional é realizado por um cardiologista especialista na análise de tacogramas.

Todas as séries temporais de intervalos RR ou tacogramas desse banco de dados são obtidas pelo monitor cardíaco Polar. Em resumo, cada tacograma foi primeiro filtrado automaticamente com auxílio do sistema próprio de filtragem do software do Polar, que utiliza um algoritmo próprio de interpolação. Sequências que possuíam uma quantidade de intervalos considerados artefatos pelo sistema Polar acima de 5% foram descartadas, salvo melhor juízo de um avaliador humano.

A seguir, os tacogramas selecionados foram submetidos à inspeção visual, realizadas por um cardiologista treinado no reconhecimento dos diversos tipos de artefatos em séries temporais eletrocardiográficas, com possíveis remoções de artefatos não eliminados anteriormente. Após essa depuração, as séries temporais de intervalos RR obtidas foram utilizadas para comparação com as séries equivalentes construídas pelo método da filtragem adaptativa.

3.4.2 Filtragem adaptativa

A filtragem adaptativa é usada como um procedimento automático para filtrar as séries de intervalos RR. O método utilizado é o apresentado em (WESSEL et al., 2000) levando-se em conta a natureza dos dados analisados. A filtragem adaptativa pode ser aplicada com diferentes finalidades, por exemplo, identificação de sistemas, inversão de sistemas, predição de sinal e cancelamento de interferências (HAYKIN, 1996). Nesse contexto o método de filtragem adaptativo foi aplicado com a finalidade de cancelamento de interferências, uma vez que, as séries temporais de intervalos RR podem sofrer interferências de artefatos, oriundos de diversas fontes, mau contato de eletrodo e outros, que não caracterizam o ritmo sinusal presente no dado para análise da VFC.

Para cancelamento de interferências, esse método pode ser usado de duas maneiras: como proposto em Wessel et al. (2000) ou simplesmente na identificação dos inter-

valos RR não caracterizados como ritmo sinusal normal pelo filtro e sua extração da série. Para ambas as situações, o algoritmo da filtragem adaptativa usado é assim denominado pois é baseado na média e desvio padrão que mudam ao longo da série analisada conforme a variabilidade apresentada.

Esse filtro consiste em três procedimentos: (a) remoção de intervalos RR menores que 350ms, sem limite superior para o conjunto de dados analisados nesse trabalho, (b) procedimento adaptativo e (c) procedimento adaptativo de controle.

A primeira etapa remove os intervalos RR menores que 350ms, uma vez que, podem equivaler ao período refratário absoluto ou relativo (onde fisiologicamente não ocorre novo batimento cardíaco) ou equivocadamente representar aumento da freqüência cardíaca ultrapassando o número de batimentos por minuto (bpm) compatíveis com o ritmo sinusal em termos clínicos humanos.

O procedimento adaptativo calcula a média e desvio padrão adaptativo. As etapas nesse procedimento são:

(i) estimar a variabilidade básica da série utilizando-se uma série binomial. Dado um tacograma $x_1, x_2, ..., x_N$, sendo N o número de intervalos RR da série, uma nova série 6-binomial é dada por:

$$t_i = \frac{x_{i-3} + 6x_{i-2} + 15x_{i-1} + 20x_i + 15x_{i+1} + 6x_{i+2} + x_{i+3}}{64} \tag{3.1}$$

(ii) média (μ) (aqui $\mu_1 = \bar{X}$, onde $\bar{X}$ é a média de $x_1, x_2, ..., x_N$) e o desvio padrão adaptativo (σ) em uma nova série é escrita como

$$\mu_i = \mu_{i-1} - c(\mu_{i-1} - t_{i-1}) \tag{3.2}$$

$$\sigma_i = \sqrt{\mu_i{}^2 - \lambda_i} \tag{3.3}$$

sendo c o coeficiente de controle, $c \in [0, 1]$ (aqui $c = 0,05$), e λ_i (aqui $\lambda_1 = \mu_1^2$) é o segundo momento adaptativo:

$$\lambda_i = \lambda_{i-1} - c(\lambda_{i-1} - t_{i-1}^2) \tag{3.4}$$

O coeficiente de controle c deve ser escolhido apropriamente, pois causa grande

impacto na geração da média e do desvio padrão adaptativo. De fato, pode ser escrito como sendo recorrente, conforme a Equação 3.2 na forma matricial:

$$\begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \vdots \\ \mu_N \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1-c & c & 0 & \cdots & 0 \\ (1-c)^2 & c(1-c) & c & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ (1-c)^{N-1} & c(1-c)^{N-2} & c(1-c)^{N-3} & \cdots & c \end{bmatrix} \times \begin{bmatrix} t_0 \\ t_1 \\ t_2 \\ \vdots \\ t_N - 1 \end{bmatrix} \quad (3.5)$$

onde, na coluna da direita, $t_0 = \bar{X}$ é a média da série original $X = \{x_1, x_2, ..., x_N\}$. Note que cada coluna na matriz $N \times N$ soma-se $[c_{ij}]$ à unidade, sendo $\sum_{j=1}^{N} c_{ij} = 1$ e $\mu_i = \sum_{j=1}^{i} c_{ij} t_{j-1}$, vemos que μ_i é o valor médio da amostra $\{t_0, t_1, ..., t_{i-1}\}$ com fatores de ponderação c_{ij}.

Logo, reformulando as Equações 3.3 e 3.4 como $\sigma_i^2 = \mu_i^2 - \sum_{j=1}^{i} c_{ij} t_{j-1}^2$, tem-se que σ_i é identificado como o desvio padrão de $\{t_0, t_1, ..., t_{j-1}\}$.

(iii) a regra de exclusão (que irá remover o intervalo RR analisado), considerando o intervalo RR x_N como não normal, é dada por:

$$\begin{aligned} |x_i - x_{i-1}| &> \frac{\rho}{100} x_{i-1} + a * \bar{\sigma} \quad \text{e} \\ |x_i - x_v| &> \frac{\rho}{100} x_v + a * \bar{\sigma} \end{aligned} \quad (3.6)$$

onde ρ equivale a um limite proporcional e $a*\bar{\sigma}$ é a regra generalizada $3-$sigmas (considera que até três desvios padrões da média ainda pertence a mesma distribuição), $\bar{\sigma}$ é a média de σ e x_v é o último intervalo RR válido. Os valores ditos como não normais são substituídos por um valor aleatório entre $[\mu_i - \frac{1}{2}\sigma_i, \mu_i + \frac{1}{2}\sigma_i]$ evitando assim, falso decréscimo na variabilidade.

O procedimento adaptativo de controle é usado como precaução, onde a série formada após as possíveis substituições, $x'_1, x'_2, ..., x'_N$, passa novamente por todo processo descrito acima acrescendo mais uma regra de exclusão, dada por:

$$|x'_i - \mu_i| > a * \sigma_i + \sigma b \quad (3.7)$$

sendo σb representa a variabilidade básica introduzida para reduzir a quantidade de possíveis artefatos com o filtro considerando séries temporais de intervalos RR

com baixa variabilidade. Os intervalos RR que são considerados não normais, segundo essa regra de exclusão, são substituídos pelo valor correspondente na série $t_0, t_2, ..., t_{N-1}$.

Considerando o exemplo mostrado na Figura 3.5 temos a mesma série filtrada com diferentes valores de c. Para os diferentes valores de c usados, a Figura 3.6 mostra como μ e σ adaptativos (Equações 3.2 e 3.3) variam conforme o índice i, onde os tacogramas foram artificialmente deslocados verticalmente para efeito de melhor comparação.

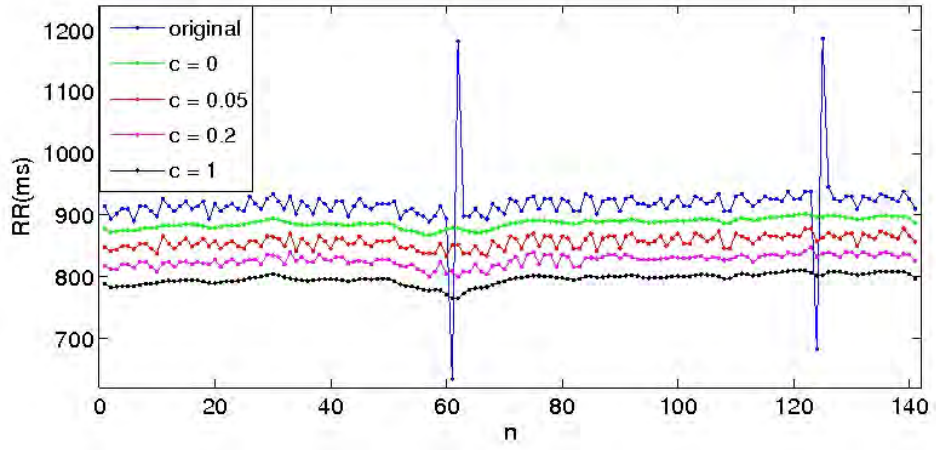

Figura 3.5 - Exemplo de um segmento de série de intervalos RR ($n = 141$) com artefatos não removidos e as séries filtradas correspondentes obtidas com diferentes valores de c. Os outros parâmetros usados são: $a = 3$, $\rho = 10$ e $\sigma_b = 0,02s$. **Para melhor visualização os tacogramas foram deslocados verticalmente**.

Observa-se na Figura 3.6 que quando o parâmetro $c = 0$ os valores de μ e σ permanecem constantes ao longo da série, com $\mu_i = t_0$ e $\sigma_i = 0$. Quando $c = 1$ a média das séries filtradas é dada por um termo simples $\mu_i = t_{i-1}$ e a variância também desaparece, isto é, $\sigma_i = 0$. Logo, valores extremos $c = 0$ ou $c = 1$ não permitem adaptação de μ e σ em alterações de variabilidade da série. Por outro lado, quando c assume valores entre 0 e 1, μ e σ se adaptam a variabilidade da série.

Para compreender como o procedimento de substituição de intervalos RR ocorre, a Figura 3.7 mostra a região do pontos $n = 121$ a $n = 129$ da série original de intervalos RR (Figura 3.5), juntamente com os gráficos para μ e σ quando o parâmetro é variado. Note que, os valores de RR consecutivos $n = 124$, 125 e 126 são filtrados.

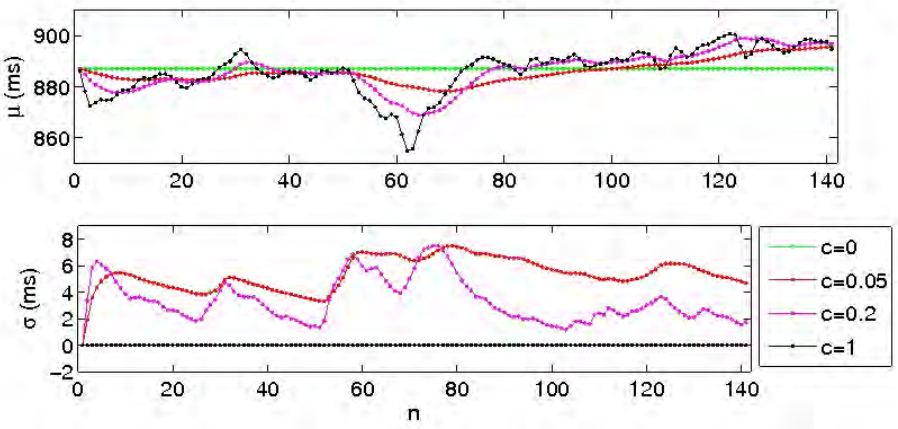

Figura 3.6 - Variação de μ e σ ao longo da série. Note que quando $c = 0$ não há adaptação da média e desvio padrão na série; quando $c = 1$ o desvio padrão também não varia.

Após a filtragem, os valores originais 652, 1156, e 916ms são substituídos por $892,9$, $893,2$, e $893,5$ms respectivamente (ver Figura 3.7).

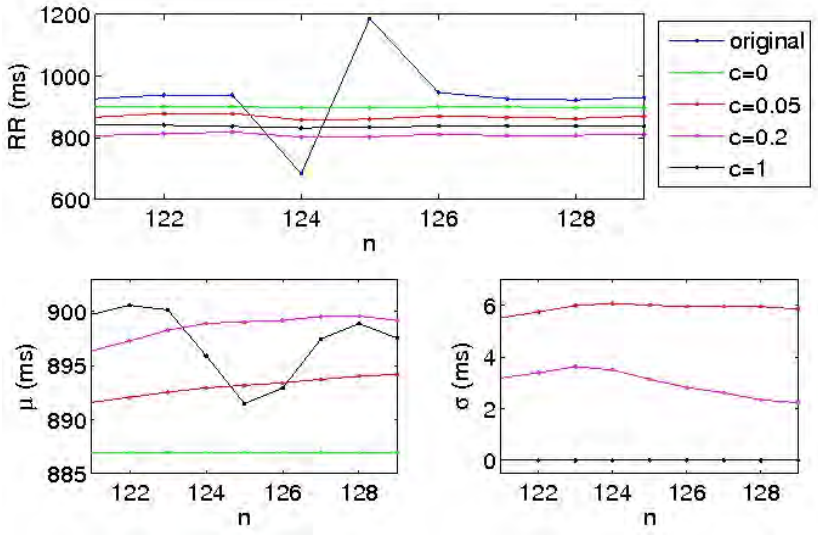

Figura 3.7 - Variação de μ e σ ao longo do pontos $n = 121 - 129$ usando diferentes valores de c. **As séries na parte superior da Figura foram deslocadas verticalmente para melhor visualização**.

Para diferentes valores de ρ, observa-se na Figura 3.8 que para $\rho = 0$, muitos intervalos RR são removidos descaracterizando o sinal original uma vez que alguns destes

pontos não seriam substituídos, de acordo com a série de filtragem realizada pelo método convencional (especialista). Por outro lado, quando $\rho = 30$ não substitui pontos que seriam removidos.

A variação de μ e σ ao longo das séries é mostrada na Figura 3.9 para diferentes valores de ρ. A maior variação em μ e σ corresponde a $\rho = 30$, evidenciando uma grande diferença entre os intervalos RR.

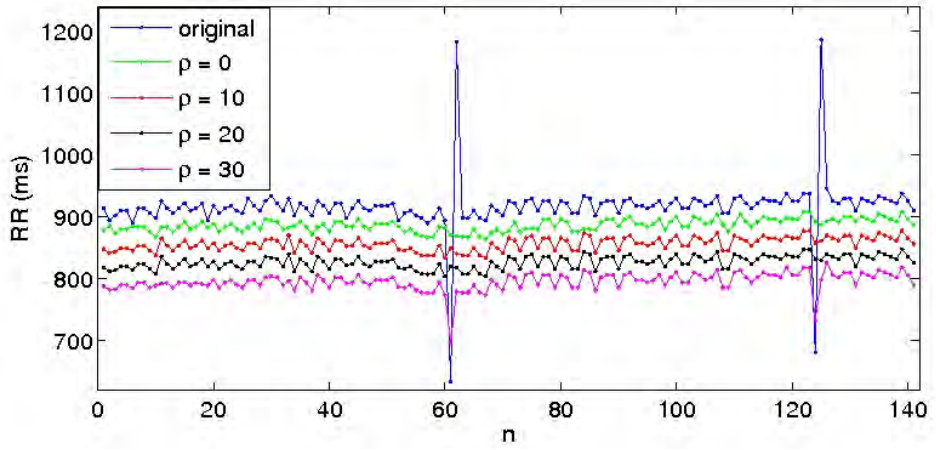

Figura 3.8 - Exemplo de um segmento de série ($n = 141$ intervalos RR) com artefatos não removidos e as séries filtradas correspondente com diferentes valores de ρ. Os demais parâmetros usados são: $a = 3$, $c = 0,05$ e $\sigma_b = 0,02s$. Para melhor visualização os tacogramas são deslocados verticalmente.

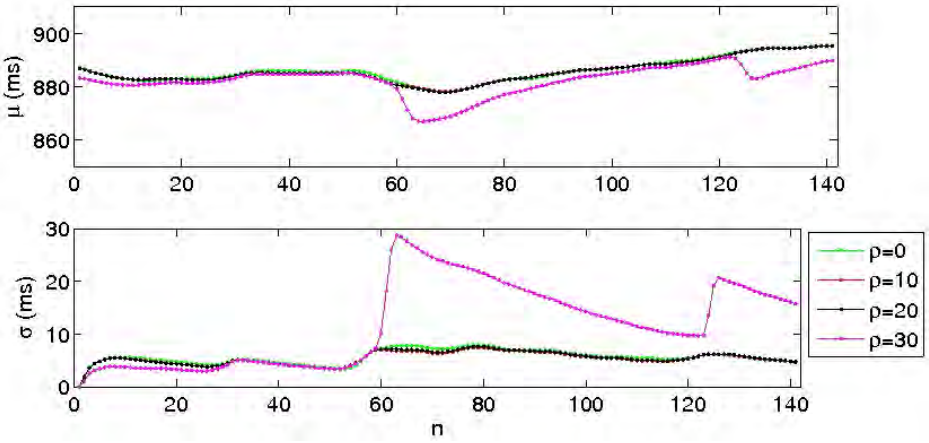

Figura 3.9 - Variação de μ e σ ao longo da série com diferentes valores de ρ.

A importância de aplicação do filtro adaptativo nesse conjunto de dados está relacionada a automatização do pré-processamento das séries temporais, facilitando assim a análise do especialista. Outro aspecto importante nessa aplicação é o ajuste dos parâmetros livres do filtro que devem ser escolhidos adequadamente para que somente os pontos considerados artefatos (tanto pelo filtro como pela análise do especialista) sejam removidos da série.

O filtro pode ser usado de duas formas: a primeira conforme descrito acima e na segunda forma o filtro pode ser usado como identificador dos intervalos RR que, pela definição do filtro adaptativo, são considerados não normais.

3.4.3 Duas maneiras de utilização do filtro

Resumidamente o filtro adaptativo identifica, conforme o ajuste de seus parâmetros, os intervalos RR considerados não normais e realiza (de maneira aleatória) a substituição destes intervalos. Para substituição desses intervalos, o filtro seleciona um valor aleatório entre $\pm\mu + (1/2\sigma)$. Entretanto, esse mesmo procedimento adaptativo de "encontrar" os intervalos RR não normais pode ser usado de duas maneiras. A primeira conforme já descrito e a segunda apenas como identificador dos pontos não normais que serão removidos da série, sem serem substituídos.

Essa segunda maneira de utilização pode interferir na sequência temporal da série, mas por outro lado também pode garantir que todos os intervalos RR restantes correspondem aos critérios anteriormente estabelecidos (de "normalidade"). O processo de remoção dos pontos da série temporal deve ser cuidadoso para não realizar alterações indesejadas na análise da dinâmica das séries temporais. Para tal, a remoção dos pontos deve ser diferenciada e de acordo com o método de dinâmica não linear utilizada na análise dos sinais.

A Tabela 3.1 ilustra o procedimento de retirada de intervalos RR ditos não normais, usando como identificador o filtro adaptativo, conforme a ferramenta utilizada na análise. Observe que nesta Tabela 3.1 estão ilustradas a série temporal (RR), a série de diferenças dos intervalos RR (ΔRR), o gráfico de diferenças de segunda ordem (CTM) e o cálculo da entropia de Shannon a partir das séries de símbolos (dinâmica simbólica).

Para as séries de diferenças sucessivas entre os intervalos RR (ΔRR) são removidas também as diferenças influenciadas pelo RR removido, por exemplo, se x_i é considerado não sinusal pelo algoritmo, as coordenadas (χ_i), (χ_{i-1}) e (χ_{i-2}) são removidas da análise também.

Já para o cálculo da entropia de Shannon, todas as palavras que poderiam conter um ponto não normal são removidas da análise. Logo, se x_i é considerado não sinusal pelo algoritmo, são removidos os pontos no intervalo $[w_{i-l}, w_i]$. Os pontos que permanecem na série são somente aqueles que o algoritmo considera como ritmo sinusal e que não estariam na mesma palavra daqueles pontos removidos.

Um problema associado a este tipo de utilização do filtro é que a quantidade de intervalos RR que são removidos de cada série é diferenciada de acordo com a necessidade de cada método de análise utilizado. Isso exige um cuidado redobrado no

Tabela 3.1 - Exemplo de remoção do intervalo RR detectado pelo filtro da análise. ΔRR é a diferença dos intervalos RR ($x_{i+1} - x_i$), CTM é a medida da tendência central e Shannon o valor da entropia para as séries de dinâmica simbólica. No exemplo abaixo, o intervalo RR detectado (em vermelho) é x_7 e os demais são aqueles que serão removidos da análise. Observe que, conforme a ferramenta utilizada, a quantia de pontos removidos é diferente.

RR	Δ RR	CTM	Shannon	
x_1	$\delta RR_1 = x_2 - x_1$	$\chi_1 = (\delta RR_1, \delta RR_2)$	s_1	$w_1 = (s_1, s_2, s_3, s_4, s_5)$
x_2	$\delta RR_2 = x_3 - x_2$	$\chi_2 = (\delta RR_2, \delta RR_3)$	s_2	$w_2 = (s_2, s_3, s_4, s_5, s_6)$
x_3	$\delta RR_3 = x_4 - x_3$	$\chi_3 = (\delta RR_3, \delta RR_4)$	s_3	$w_3 = (s_3, s_4, s_5, s_6, s_7)$
x_4	$\delta RR_4 = x_5 - x_4$	$\chi_4 = (\delta RR_4, \delta RR_5)$	s_4	$w_4 = (s_4, s_5, s_6, s_7, s_8)$
x_5	$\delta RR_5 = x_6 - x_5$	$\chi_5 = (\delta RR_5, \delta RR_6)$	s_5	$w_5 = (s_5, s_6, s_7, s_8, s_9)$
x_6	$\delta RR_6 = x_7 - x_6$	$\chi_6 = (\delta RR_6, \delta RR_7)$	s_6	$w_6 = (s_6, s_7, s_8, s_9, s_{10})$
x_7	$\delta RR_7 = x_8 - x_7$	$\chi_7 = (\delta RR_7, \delta RR_8)$	s_7	$w_7 = (s_7, s_8, s_9, s_{10}, s_{11})$
x_8	$\delta RR_8 = x_9 - x_8$	$\chi_8 = (\delta RR_8, \delta RR_9)$	s_8	$w_8 = (s_8, s_9, s_{10}, s_{11}, s_{12})$
⋮	⋮	⋮	⋮	⋮

pré-processamento das séries de intervalos RR. Considerando todo o nosso conjunto de dados, esse procedimento é avaliado apenas em séries longas, tais como as de duração de 24 horas (Physionet e Coria).

Considerações sobre o Capítulo

Esse Capítulo descreveu o estudo de caso sobre as séries de intervalos RR, reforçando sua importância e relevância usando a metodologia proposta. O pré-processamento adotado para esse conjunto de dados foi a filtragem. Foram apresentados os dois tipos de filtragem: a convencional realizada pelo especialista e dita como o "padrão ouro" e a adaptativa que foi proposta neste estudo.

É importante salientar que os parâmetros da filtragem adaptativa são ajustados de forma a se aproximarem do procedimento convencional e facilitar o pré-processamento por um especialista. Uma vez que, esse tipo de pré-processamento é demorado, demandando muito tempo e expertise de um especialista. No Capítulo 4 serão apresentados os resultados da análise dos dados e pré-processamento desse estudo de caso, juntamente com os resultados obtidos da metodologia proposta.

4 RESULTADOS: PRIMEIRO ESTUDO DE CASO

Esse capítulo de resultados obtidos do primeiro estudo de caso está dividido em duas grandes Seções (4.1 e 4.2), sendo que o objetivo principal deste trabalho é discriminar os conjuntos de séries temporais usando diferentes métodos não lineares e mineração de dados. A Seção 4.1 descreve a análise a partir dos conjuntos de séries de intervalos RR, onde o pré-processamento é avaliado e a variabilidade da frequência cardíaca é analisada com as ferramentas de sistemas dinâmicos. Nessa Seção são apresentadas as contribuições originais do trabalho envolvendo o ajuste e a análise do uso do filtro adaptativo na séries de intervalos RR.

Na Seção 4.2 são apresentados os resultados com a metodologia proposta usando as saídas obtidas dos classificadores J48 e SVM. A partir desses resultados caracterizamos a dinâmica das séries temporais pelas medidas dos métodos de dinâmica linear e não linear usados. Nessa Seção são apresentadas as contribuições originais do trabalho sobre a capacidade dos classificadores em detectar diferenças entre as dinâmicas dos grupos de tacogramas.

4.1 Análise dos dados

Os resultados da análise dos conjuntos de séries temporais de intervalos RR são apresentados em subseções distintas, conforme o objetivo de cada análise. Entre os objetivos dessa análise estão: analisar o método de filtragem proposto, comparando com o método de filtragem convencional; verificar o impacto da extensão das séries de intervalos RR para análise de VFC e a influência dessa extensão para predição de eventos clínicos adversos (eventos extremos, por exemplo, a morte do paciente).

Para todas as análises citadas são usados os parâmetros do gráfico de Poincaré (BRENNAN et al., 2001) como medidas para as estatísticas realizadas. Os parâmetros do gráfico de Poincaré estão diretamente relacionados à fisiologia do coração e ao sistema nervoso autônomo.

O parâmetro SD1 está relacionado aos intervalos RR de curto alcance, mostrando a variabilidade dos sucessivos intervalos, ligado ao controle parassimpático do nodo sinusal, enquanto SD2 está relacionado aos intervalos RR de longo alcance, evidenciando a VFC ao longo da medida obtida, ligado ao controle simpático do nodo sinusal pelo sistema nervoso autônomo (MOUROT et al., 2004). Alguns trabalhos mostram a análise dessas variáveis em VFC para compreensão dos fenômenos fisiológicos e biológicos ocorridos durante treinamento físico (MOUROT et al., 2004), em

pacientes com insuficiência cardíaca congestiva (ISLER, 2007) e durante a prática da meditação (GOSHVARPOUR et al., 2011).

Eventualmente, conjuntos de séries de intervalos RR diferentes são usados em objetivos distintos, devidamente justificados.

4.1.1 Filtragem convencional e a filtragem adaptativa

Neste estudo, o objetivo principal foi comparar os resultados obtidos com as variáveis do gráfico de Poincaré (SD1, SD2 e SD1/SD2) das séries filtradas a partir dos dois métodos de filtragem: adaptativo e convencional. O método adaptativo foi usado com a finalidade de automatizar o processo realizado por um especialista. Esse método serve como uma alternativa facilitando a filtragem.

Os parâmetros do filtro adaptativo (ρ e c) foram ajustados empiricamente usando como referência as séries filtradas pelo especialista. É importante salientar que não descartamos a filtragem realizada por um especialista, onde seu conhecimento foi usado para análise das séries filtradas pelo método adaptativo, contribuindo para o ajuste dos parâmetros do filtro. Ou seja, o cardiologista realiza uma análise visual nas séries filtradas verificando a equivalência com sua confiabilidade clínica. A partir desse passo, o processo é automatizado, não havendo mais interferência no ajuste dos parâmetros ou necessidade de participação do cardiologista.

Para comparação entre os dois métodos de filtragem, foi usado um conjunto de tacogramas do banco de dados NUTECC. A escolha em particular desse banco de dados deu-se pela possibilidade de obter as séries de intervalos RR originais e as mesmas filtradas pelo cardiologista.

O conjunto de dados foi composto de 229 tacogramas: 53 de recém nascidos prematuros hospitalizados em unidade de tratamento intensivo (UTI) (grupo G1), 12 de recém nascidos normais (grupo G2), 62 de adultos jovens saudáveis (grupo G3), 41 adultos submetidos a dieta de baixa caloria para perda de peso (grupo G4) e 61 adultos em avaliação pré-operatória para cirurgia de revascularização do miorcárdio (grupo G5).

O procedimento de filtragem permite o reconhecimento e substituição (se for o caso) de intervalos RR que correspondem a artefatos. Esses artefatos podem ocorrer nas séries devido a interferência eletrônica, movimentos do paciente e deslocamento de eletrodos durante a captação do sinal pelo equipamento.

A Figura 4.1 mostra dois exemplos de séries temporais de intervalos RR que podem conter artefatos. A Figura 4.1A com a presença evidente de artefatos, representados pelos picos de maior amplitude e a 4.1B sem a presença evidente de artefatos (podendo existir artefatos não evidentes visualmente).

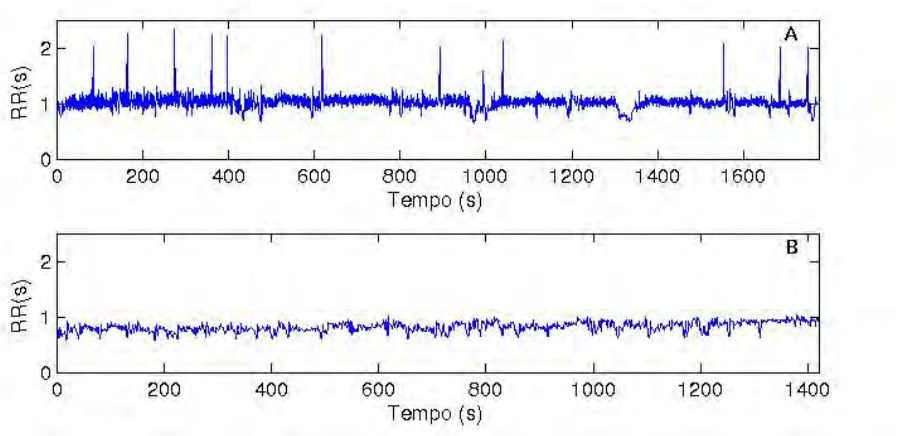

Figura 4.1 - Dois tacogramas com mais de 1400 intervalos RR cada. a) Exemplo de tacograma mostrando visíveis artefatos e b) tacograma sem visíveis artefatos.

Analisando as Figura 4.2 e 4.4 não foi possível estabelecer diferenças visuais comparando as séries filtradas com os dois métodos a partir da série sem filtragem captada.

A Figura 4.2 apresenta uma série temporal de intervalos RR de um adulto jovem saudável contendo 2700 intervalos. A porcentagem de artefatos[1] estimada pelo método adaptativo foi de $0,40\%$ e pelo método convencional foi de $2,51\%$. Comparando visualmente as duas séries filtradas não foi possível estabelecer diferenças nítidas. Analisando os respectivos gráficos de Poincaré apresentados na Figura 4.3, podemos verificar que os gráficos das séries filtradas confirmam a similaridade entre si.

A Figura 4.4 apresenta uma série temporal de intervalos RR de um recém nascido prematuro contendo 2400 intervalos. A porcentagem de artefatos estimada pelo método adaptativo foi de $7,95\%$ e pelo método convencional foi de $7,3\%$. A série original visualmente apresenta muitos picos de maior amplitude, que poderiam ser

[1] A porcentagem de artefatos é a quantidade percentual de pontos extraídos da série dada pela etapa 1 no método adaptativo e pela quantia de pontos extraídos pelo especialista no método convencional.

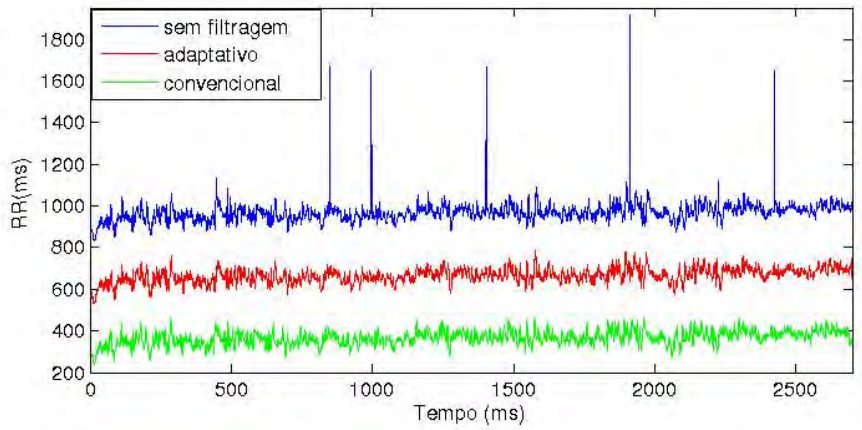

Figura 4.2 - Tacograma de um adulto jovem saudável (diagnóstico médico) contendo 2700 intervalos RR. A porcentagem de artefatos estimado usando o método adaptativo foi de $0,40\%$ e usando o método convencional foi de $2,51\%$. **Para melhor visualização os tacogramas foram deslocados verticalmente**.

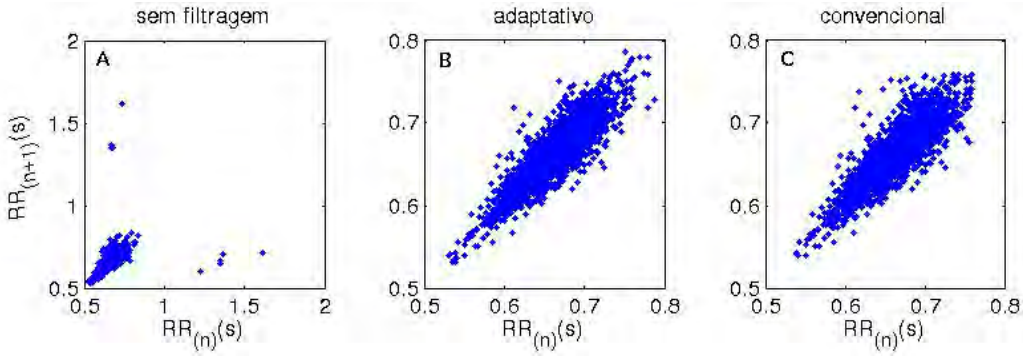

Figura 4.3 - Gráficos de Poincaré para os tacogramas da Figura 4.2. a) Série original, b) série filtrada pelo método adaptativo e c) série filtrada pelo método convencional.

identificados como artefatos, segundo os ajustes realizados no filtro. Para analisar as distinções entre as duas séries filtradas, a Figura 4.5 apresenta os gráficos de Poincaré. Foi possível estabelecer diferenças entre os dois gráficos, entretanto, a distribuição dos pontos foi similar.

Para analisar a diferença entre os dois métodos de filtragem, foram calculados os parâmetros SD1, SD2 e SD1/SD2 do gráfico de Poincaré para cada série filtrada. E em seguida, para análise comparativa estatística, o teste t de Student não pareado (em caso de distribuição não gaussiana foi realizado o teste de Mann-Whitney) (valor p) (mais detalhes sobre os testes estatísticos ver Apêndice A).

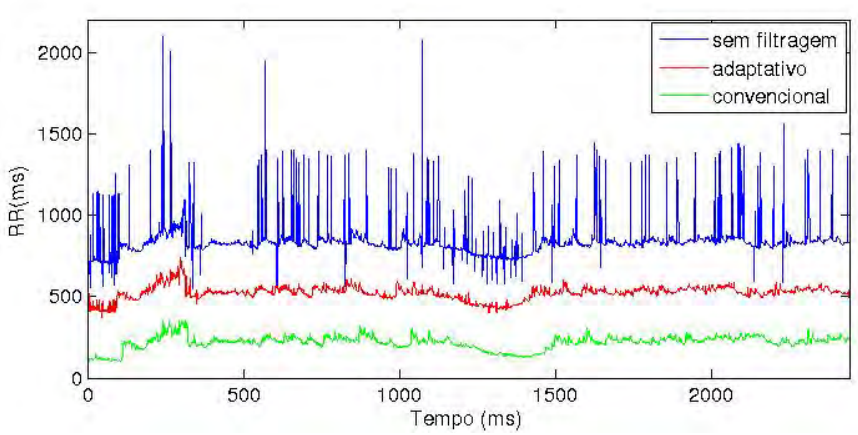

Figura 4.4 - Tacograma de um recém nascido prematuro contendo 2440 intervalos RR. A porcentagem de artefatos usando o método adaptativo foi de $7,95\%$ e usando o método convencional foi $7,3\%$. **Para melhor visualização os tacogramas foram deslocados verticalmente**.

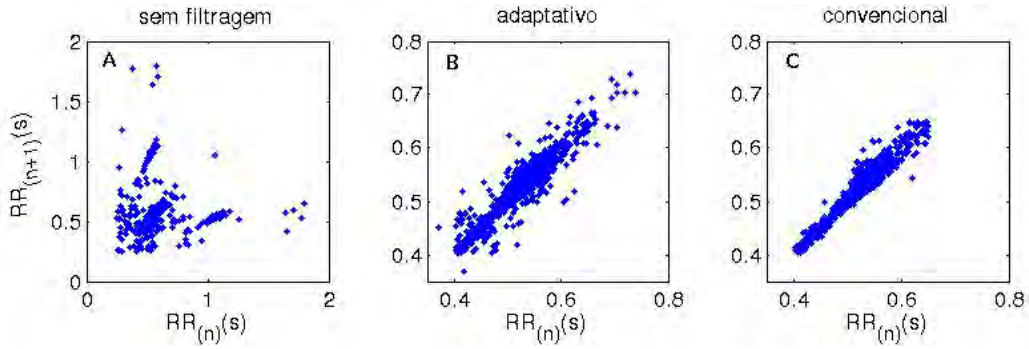

Figura 4.5 - Gráficos de Poincaré para os tacogramas da Figura 4.4. a) Série original, b) série filtrada pelo método adaptativo e c) série filtrada pelo método convencional.

Foi admitido erro alfa de 5%, considerando-se que as variáveis são significativamente diferentes para valores de $p < 0,05$. Para validação do método proposto em relação ao método convencional, foi aplicado o coeficiente de correlação de Pearson (r) (ou o de Spearman, conforme a distribuição dos dados) (ver Apêndice A). Esses coeficientes medem o grau de correlação entre duas variáveis.

O valor obtido para r deve estar entre -1 (correlação máxima negativa) e $+1$ (correlação máxima positiva). Se o valor r estiver entre 0 e $0,3$ (positivo ou negativo) indica uma fraca correlação, entre $0,3$ e $0,7$ (positivo ou negativo) indica uma correlação moderada e entre $0,7$ e 1 (positivo ou negativo) indica uma forte correlação.

A Tabela 4.1 apresenta a análise comparativa dos resultados obtidos com os filtros

adaptativo e convencional usando os valores dos parâmetros do gráfico de Poincaré. Observamos que os valores de correlação, com exceção do parâmetro SD1/SD2 para o grupo G1, são maiores que $0,7$ evidenciando uma forte correlação entre os dois métodos de filtragem.

Em relação ao valor p, quase todas o casos comparados são estatisticamente semelhantes entre as médias, exceto o grupo G1 com respeito às variáveis SD1 e SD1/SD2. A distribuição dos parâmetros de Poincaré é apresentada na Figura 4.6, onde é enfatizado que o grupo G1 é mais suscetível a produção de artefatos, logo pode-se dizer que, mais difícil de medir VFC de recém nascidos prematuros que nos demais.

Tabela 4.1 - Análise comparativa dos resultados obtidos com os filtros adaptativo e convencional com referência aos valores das variáveis SD1, SD2 e SD1/SD2 correspondendo a cinco diferentes situaçãoes clínicas. O coeficiente de correlação de Pearson ou Spearman (r) maior que $0,7$ indica uma forte correlação. Valores de $p < 0,05$ no teste t de Student não pareado ou Mann-Whitney indica que a média dos valores das variáveis em cada grupo são significativamente diferentes.

	Grupo	Adaptativo	Convencional	r	valor p
SD1	G1	$7,4268 \pm 3,2704$	$5,5171 \pm 3,0612$	$0,8463$	$0,0022^{S}$
	G2	$8,2613 \pm 3,6262$	$6,8326 \pm 2,9826$	$0,7075$	$0,3033$
	G3	$33,8593 \pm 14,3990$	$37,1512 \pm 17,4353$	$0,9794$	$0,2578$
	G4	$17,6689 \pm 10,5331$	$19,1482 \pm 12,2198$	$0,9753$	$0,5400$
	G5	$14,3142 \pm 7,8036$	$16,3038 \pm 10,4709$	$0,8252^{*}$	$0,3785^{**}$
SD2	G1	$37,9816 \pm 19,0262$	$36,2147 \pm 15,6474$	$0,7380$	$0,5872$
	G2	$50,4241 \pm 15,3866$	$48,9572 \pm 16,0971$	$0,9822$	$0,8216$
	G3	$85,5807 \pm 29,9954$	$85,4594 \pm 30,1717$	$0,9813$	$0,9823$
	G4	$70,0338 \pm 34,2140$	$67,3651 \pm 36,0056$	$0,9339$	$0,7323$
	G5	$47,8016 \pm 22,8041$	$48,5859 \pm 23,3733$	$0,9868$	$0,8515$
SD1/SD2	G1	$0,2125 \pm 0,0800$	$0,1584 \pm 0,0650$	$0,5702$	$0,0002^{S}$
	G2	$0,1711 \pm 0,0819$	$0,1396 \pm 0,0389$	$0,9907^{*}$	$0,4428^{**}$
	G3	$0,3957 \pm 0,1181$	$0,4317 \pm 0,1436$	$0,9872$	$0,1325$
	G4	$0,2475 \pm 0,0890$	$0,2769 \pm 0,1041$	$0,9528$	$0,1713$
	G5	$0,3176 \pm 0,1339$	$0,3541 \pm 0,1794$	$0,9808^{*}$	$0,2927^{**}$

G1 - recém nascidos prematuros; G2 - recém nascido normal; G3 - adultos jovens saudáveis; G4 - adultos em dieta de baixa calorias; G5 - adultos com doença coronariana severa.
^{S}Significativamente diferentes; *Correlação de Spearman; **teste de Mann-Whitney.

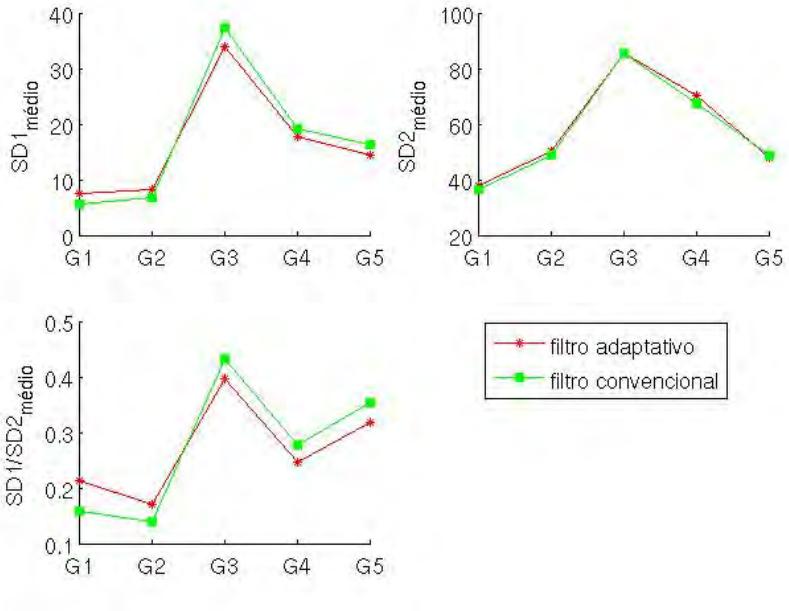

Figura 4.6 - Padrão de distribuição médio para os parâmetros do gráfico de Poincaré para os cinco grupos clínicos diferentes caracterizando os filtros adaptativo e convencional.

4.1.2 Impacto da extensão das séries temporais de intervalos RR para análise da VFC

As extensões dos tacogramas podem variar conforme o tempo de aquisição e condições específicas do indivíduo. Assim, é mais difícil captar sinais de um recém nascido prematuro (SELIG et al., 2011) do que em adultos. Por outro lado, espera-se que a análise de tacogramas com maiores extensões contenha mais informações, indenpendente da dificuldade de captação. Entretanto, é importante estabelecer a extensão mínima das séries temporais para representar estatisticamente a mesma série homóloga de maior tamanho.

Baseado nesse contexto, o objetivo dessa análise é verificar o impacto da extensão das séries temporais de intervalos RR na análise da VFC realizando comparações intragrupos. Nessa análise, também foram usados os valores da correlação de Pearson obtidos das variáveis do gráfico de Poincaré. Para tal, foram usados dois conjuntos de dados, um conjunto de tacogramas do PhysioNet (GOLDBERGER et al., 2000) que foi captado usando o monitor Holter (Marquette 8500 Holter) e um conjunto de tacogramas do NUTECC captado com monitor Polar.

O conjunto de dados do PhysioNet foi composto de 29 tacogramas de pacientes com falha cardíaca congestiva (CHF). As séries com duração de 24 horas foram coletadas usando o monitor Holter de pacientes entre 34 e 79 anos de idade (KRUM et al., 1995; GOLDSMITH et al., 1997). O tamanho das séries foi progressivamente aumentado contendo 250, 500, 1000, 2000, 3000, 5000, 10000, 25000, 50000, 75000 e a série completa com extensões maiores que 75000 intervalos RR. Todas as séries foram previamente filtradas com o método de filtragem adaptativo proposto.

A Figura 4.7 apresenta a matriz de correlação para cada um dos parâmetros do mapa de primeiro retorno entre os diferentes tamanhos de tacogramas usados. Observe que a variável SD1 mantém uma excelente correlação com valores superiores a $0,7$, desde o conjunto de 250 até 75000 intervalos RR. Indicando assim que esta variável pode ser comparada em diferentes estudos da literatura, mesmo se o comprimento da série temporal for diferente umas das outras. Para a variável SD2, a correlação foi superior a $0,6$, sendo o valor mínimo apresentado um pouco menor que para SD1.

Já para SD1/SD2, as correlações foram menores mostrando claramente a formação de duas regiões retangulares distintas (250 a 5000 e 5000 a $>$ 75000 intervalos). Ou seja, séries com 250 a 5000 intervalos possuem uma correlação forte com outras séries até 5000 intervalos, mas acima desse comprimento a correlação foi baixa

apresentando coeficiente em torno de $0,30$.

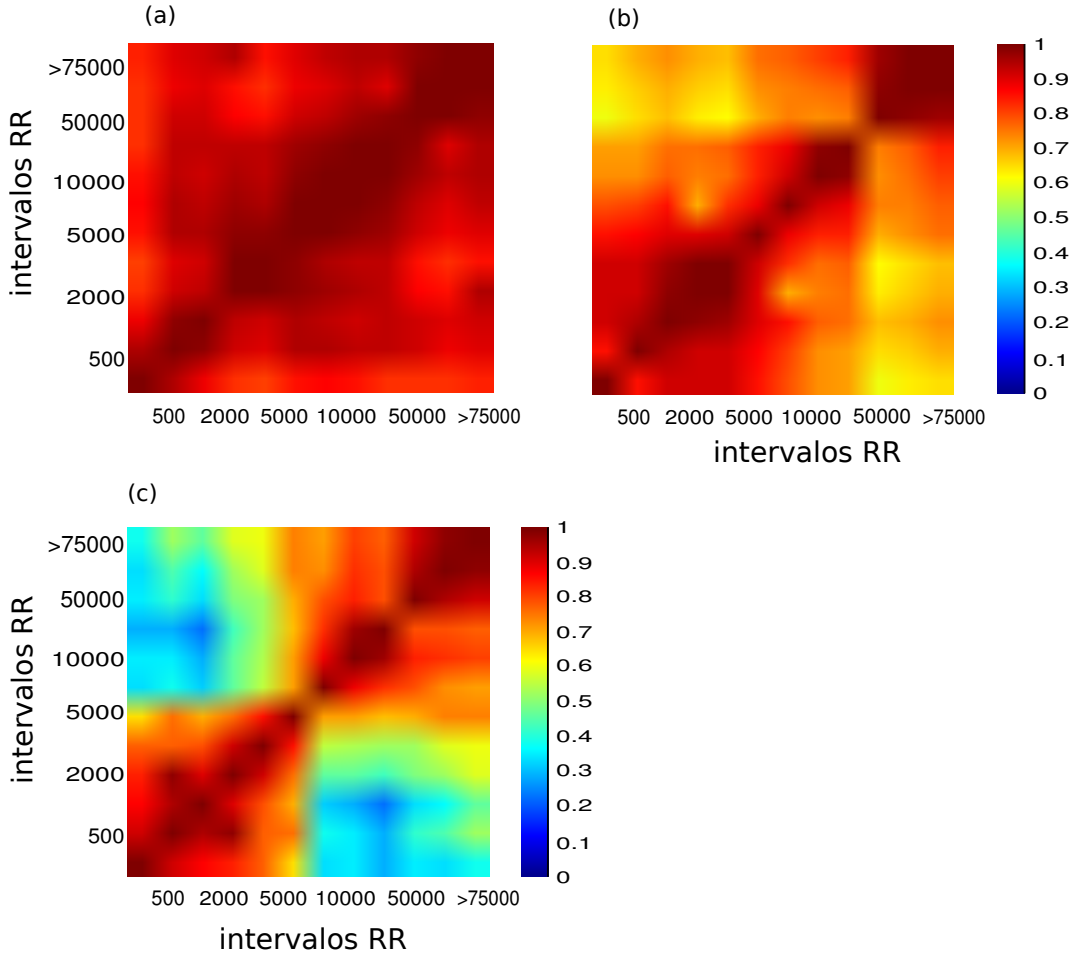

Figura 4.7 - Análise comparativa da correlação de Pearson para os parâmetros do gráfico de Poincaré para o conjunto de 29 tacogramas de CHF usando séries de intervalos RR curtas com aumento progressivo do tamanho das séries. A escala de cores à direita indica o grau de correlação. Parâmetros: (a) SD1, (b) SD2 e (c) SD1/SD2.

Para confrontar os resultados obtidos usando outra metodologia de captação dos sinais cardíacos, foram analisados tacogramas do NUTECC de indivíduos não selecionados, independentemente do sexo e que estavam em diferentes condições clínicas. Todas as séries foram captadas pelo Polar (S810i ou RS 800). No total foram 158 tacogramas sendo: 52 de recém nascidos prematuros, 20 de adultos com dieta de baixa caloria, 44 adultos jovens saudáveis, 42 adultos coronariopatas.

O comprimento de cada série temporal de intervalos RR foi progressivamente aumentado em cinco diferentes extensões para cada caso (250, 500, 1000, 2000 e o total

de intervalos coletados (entre 2000 e 3000). Inicialmente todas as séries de tamanhos totais foram filtradas para a retirada de artefatos usando-se o método de filtragem adaptativo.

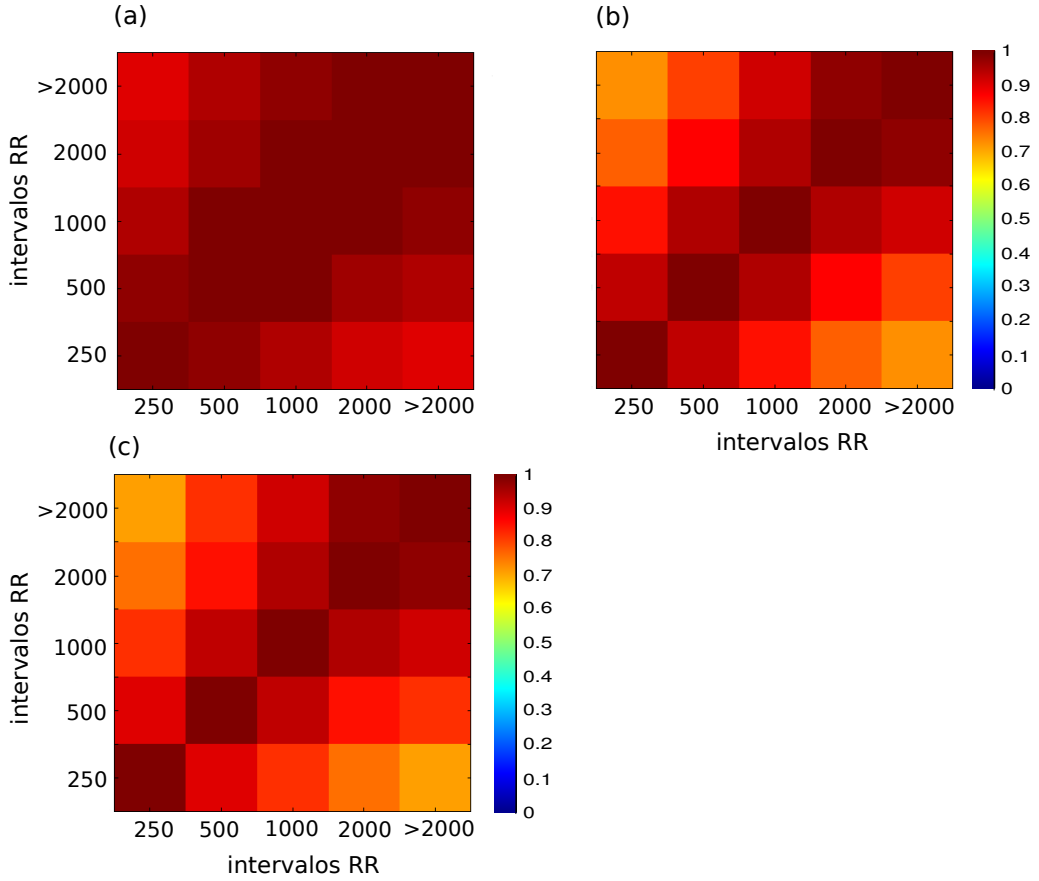

Figura 4.8 - Análise comparativa da correlação de Pearson para os parâmetros do gráfico de Poincaré para o conjunto de 158 tacogramas do NUTECC usando séries de intervalos RR curtas com aumento progressivo do tamanho das séries. A escala de cores à direita indica o grau de correlação. Parâmetros: (a) SD1, (b) SD2 e (c) SD1/SD2.

Com a finalidade de comparar as médias de cada parâmetro obtido para cada grupo de séries de tamanhos diferentes foi utilizado o teste ANOVA (ver Apêndice A). Admitiu-se erro alfa de 5% sendo que as médias são significativamente diferentes para valores de $p < 0,05$. Nos casos em que o valor p da ANOVA foi menor a $0,05$ procedeu-se ao teste de comparações múltiplas de Tukey, com comparações par a par (DRISCOLL, 1996).

A Tabela 4.2 mostra os resultados obtidos nessa análise.

Tabela 4.2 - Análise de variância (ANOVA) para as mesmas séries de intervalos divididas em cinco tamanhos (250, 500, 1000, 2000 e o tamanho máximo de cada série) para os 158 tacogramas do NUTECC, considerando os parâmetros SD1, SD2 e SD1/SD2. Valores de $p \geq 0,05$ implica na aceitação da hipótese nula (H_0), que diz que as médias não são diferentes umas das outras e $p < 0,05$ significa rejeitar H_0, ou seja, pelo menos uma das médias são diferentes do grupo analisado.

Parâmetro	Tamanhho	$\mu \pm \sigma$	ANOVA
SD1	250	$16,89 \pm 12,69$	$p = 0,81$
	500	$16,04 \pm 12,16$	
	1000	$15,52 \pm 11,86$	
	2000	$15,45 \pm 11,82$	
	Total	$15,50 \pm 11,86$	
SD2	250	$43,78 \pm 28,17$	$p = 0,003^*$
	500	$46,03 \pm 27,25$	
	1000	$48,58 \pm 26,88$	
	2000	$52,68 \pm 29,19$	
	Total	$54,74 \pm 30,21$	
SD1/SD2	250	$0,39 \pm 0,16$	$p < 0,0001^*$
	500	$0,34 \pm 0,15$	
	1000	$0,31 \pm 0,14$	
	2000	$0,29 \pm 0,13$	
	Total	$0,28 \pm 0,13$	

*Pelo menos uma das médias das séries é diferente das demais.

Para encontrar as diferenças estabelecidas pelo teste ANOVA para SD2 e SD1/SD2, a técnica de comparação múltipla de Tukey foi usada. Essa técnica compara todos os pares de média usando somente o erro alfa médio para todos os grupos. Observe na Tabela 4.3 que as séries de menor tamanho (250 e 500 intervalos) são estatisticamente diferentes daquelas com tamanho total, apesar da correlação ser alta. O menor tamanho de série que representa estatisticamente sua série homóloga foi aquela com 1000 intervalos RR.

4.1.3 Influência do comprimento das séries temporais de intervalos RR na predição de eventos clínicos adversos

Para avaliação da influência da extensão da série temporal no poder preditivo de eventos clínicos adversos, nesse caso a morte do paciente, foi utilizada especificamente a coleção das 42 séries temporais provenientes de adultos coronariopatas em pré-operatório de cirurgia de revascularização do miocárdio do banco de dados NUTECC. Optou-se por esse grupo, pois se conhecia o seguimento desses casos, tendo sido registrado que 37 apresentaram boa evolução e cinco faleceram no pós-operatório hospitalar.

Tabela 4.3 - Teste de Tukey para comparar os pares de médias que tem pelo menos uma série com diferente valor médio das demais séries (Tabela 4.2) para os parâmetros SD2 e SD1/SD2. Valores de $p \geq 0,05$ implica em aceitar H_0, ou seja, as médias não são significativamente diferentes e $p < 0,05$ signfica rejeitar H_0, as médias são diferentes significativamente.

Comparação	SD2 valor p	SD1/SD2 valor p
250 vs. 500	$p = 0,96$	$p = 0,06$
250 vs. 1000	$p = 0,56$	$p < 0,0001^*$
250 vs. 2000	$p = 0,04^*$	$p < 0,0001^*$
500 vs. 1000	$p = 0,93$	$p = 0,22$
500 vs. 2000	$p = 0,23$	$p = 0,003^*$
1000 vs. 2000	$p = 0,70$	$p = 0,53$
total vs. 250	$p = 0,006^*$	$p < 0,0001^*$
total vs. 500	$p = 0,05$	$p = 0,0004^*$
total vs. 1000	$p = 0,30$	$p = 0,25$
total vs. 2000	$p = 0,97$	$p = 0,99$

*significativamente diferentes.

Tabela 4.4 - Matriz de confusão para o parâmetro SD1 considerando o tamanho total das séries de intervalos RR dos 42 pacientes em avaliação pré-operatória.

$SD1$ - tamanho total	Morte	Não morte	Total
pacientes com resultado positivo*	5	18	23
pacientes com resultado negativo**	0	19	19
Total	5	37	42

*$SD1 \leq 11,65$ (valor de corte)
**$SD1 > 11,65$

Isso permitiu comparar o poder preditivo das variáveis do gráfico de Poincaré de acordo com o tamanho das séries temporais preparadas (250, 500, 1000, 2000 e total de intervalos RR), buscando-se o menor conjunto estatisticamente representativo do todo e que, ao mesmo tempo, mantivesse a capacidade discriminatória de ocorrência de evento clínico relevante.

Os dados foram avaliados com a realização de testes de correlação de Pearson. Estatisticamente por meio da análise de variância (ANOVA) comparando-se as médias dos valores das variáveis nos diferentes grupos. E finalmente, para comparações em termos da validação preditiva relacionada à extensão da série temporal utilizada, empregou-se a análise pela curva ROC.

A Tabela 4.4 mostra um exemplo de matriz confusão para o parâmetro SD1 considerando as séries com tamanho total dos 42 pacientes em avaliação pré-operatória, o valor de corte foi $11,65$.

A curva ROC foi usada nesse contexto para verificar o desempenho de diagnóstico (predição) estudando a sensibilidade e a especificidade dos parâmetros do gráfico de Poincaré dado um determinado valor de corte. A Tabela 4.5 mostra os valores de corte, sensibilidade (S), especificidade (E), razão positiva de verossimilhança (LRP), acurácia (A) e a área sob a curva ROC dos parâmetros do gráfico de Poincaré considerando diferentes tamanhos das séries de intervalos RR dos 42 tacogramas divididos na presença do evento analisado (5 com óbito após cirurgia e 37 não óbito após cirurgia).

Tabela 4.5 - Análise ROC para os parâmetros do gráfico de Poincaré considerando três diferentes tamanhos de séries de intervalos RR (1000, 2000 e total). Observe que, independente do tamanho das séries, os valores de S, E obtidos são similares entre si.

Parâmetro	Tamanho	Corte	S	E	A	LRP	Área sob a curva ROC
SD1	total	11,65	1	0,51	0,57	2,06	0,75 (0,46 to 1)
	1000	12,13	1	0,51	0,57	2,06	0,76 (0,47 to 1)
	2000	11,63	1	0,51	0,57	2,06	0,75 (0,46 to 1)
SD2	total	36,17	0,8	0,57	0,60	1,85	0,61 (0,30 to 0,92)
	1000	45,48	1	0,46	0,52	1,85	0,67 (0,37 to 0,97)
	2000	36,55	0,8	0,57	0,60	1,85	0,64 (0,34 to 0,95)
SD1/SD2	total	0,20	0,6	0,81	0,79	3,17	0,61 (0,30 to 0,91)
	1000	0,21	0,6	0,84	0,81	3,70	0,66 (0,36 to 0,96)
	2000	0,20	0,6	0,81	0,79	3,17	0,62 (0,31 to 0,92)

*Estimativa de Wilcoxon (intervalo de confiança de 95%).

Os diferentes tamanhos de tacogramas analisados apresentam sensibilidade e especificidade muito similares entre si, com uma área sob a curva ROC acima de 0,60 comprovando que os parâmetros mantêm um bom valor preditivo independente do tamanho da amostra, com níveis relativamente elevados de acurácia (A), de uma maneira geral.

4.2 Caracterização das séries de intervalos RR usando J48 e SVM

A metodologia proposta nesse trabalho envolve técnicas de sistemas dinâmicos para obter valores das medidas dos tacogramas e técnicas de mineração de dados para discriminar as séries conforme as medidas apresentadas aos classificadores. Foram usados dois classificadores, o J48 (usamos a implementação disponibilizada no *software* WEKA (WITTEN; FRANK, 2005)) e o SVM (usamos o pacote LIBSVM (NATIONAL TAIWAN UNIVERSITY, 2012)) que fornecem informações complementares sobre as medidas que mais caracterizam os grupos de séries de intervalos RR analisados (capazes de encontrar diferenças ou não na dinâmica entre os grupos comparados).

Ao total foram obtidas 26 medidas dos tacogramas, conforme a Tabela 4.6. As medidas listadas de 1 a 5 foram obtidas do gráfico de Poincaré, as listadas de 6 a 13 do CTM com diferentes valores de raio (ρ), de 14 é da dinâmica simbólica (DS), de 15 a 17 são da medida de complexidade, 18 do coeficiente AR e as listadas de 19 a 26 são referentes ao RQA.

Para o conjunto de séries temporais filtradas usando o filtro adaptativo ou usando o filtro convencional foram extraídas 26 medidas, conforme apresentado na Tabela 4.6. E para o conjunto de séries temporais onde foi usado o filtro como identificador de intervalos RR não normais foram extraídas 17 medidas, sendo as medidas 1 a 17 listadas na Tabela 4.6.

Em cada análise realizada com os classificadores, foram usadas apenas duas classes ou dois grupos de pacientes com diagnósticos diferentes pré estabelecidos por um médico especialista. Optou-se por um mesmo número de casos de cada classe para compor os conjuntos de treinamento e de teste, uniformizando assim a comparação entre os dois grupos.

A Tabela 4.7 apresenta todas as comparações de grupos e o número de casos usados para os conjuntos de treinamento e teste para os dois classificadores. É importante salientar que o número de casos para treinamento e teste foi estabelecido empiricamente. Foram estabelecidos três formas diferentes de pré-processamento de dados, indicados na primeira coluna da Tabela 4.7: FC - filtro convencional, FA - filtro adaptativo e PR - pontos removidos (conforme abordado na Seção 3.4.3).

Para cada comparação entre dois grupos distintos, o classificador foi executado 100 vezes, sendo que para cada nova execução um novo conjunto de treinamento e de teste foi estabelecido, mantendo-se o número de casos já determinado. O resultado

Tabela 4.6 - Medidas de sistemas dinâmicos usadas como entrada nas técnicas de MD.

Método	Nº	Índices
Gráfico de Poincaré	1	sdsd
	2	sd1
	3	sdnn
	4	sd2
	5	sd1/sd2
CTM (ρ)	6	ctm(0,01s)
	7	ctm(0,03s)
	8	ctm(0,05s)
	9	ctm(0,07s)
	10	ctm(0,09s)
	11	ctm(0,11s)
	12	ctm(0,13s)
	13	area-br
Dinâmica Simbólica	14	entr-DS
Medida de Complexidade (LMC)	15	lmc(0.25)
	16	lmc(0.5)
	17	lmc(1)
Modelo AR	18	ar
Medidas RQA	19	rr
	20	det
	21	L
	22	Lmax
	23	entr
	24	Lam
	25	tt
	26	Vmax

Observações: os diferentes índices do CTM foram obtidos variando-se o raio ρ (Subseção 2.2.3). O índice entr-DS corresponde ao valor de H na Subseção 2.2.4. Os índices LMC foram obtidos calculando-se Γ variando-se os valores de β (Subseção 2.2.5).O índice *ar* é obtido calculando EQM (Subseção 2.2.1).

do grupo de teste fornecido pelo classificador, ou seja, a capacidade de acertar a qual grupo pertencia determinado conjunto de medidas foi denominado de *acurácia*. A acurácia foi definida como a quantia de acertos no grupo de teste dividida pelo número total de casos de teste obtendo-se a acurácia média no final de todas as execuções para cada comparação entre dois grupos.

4.2.1 J48

Na análise com o classificador J48 as medidas foram apresentadas ao classificador, que após 100 execuções, obtivemos a acurácia média estabelecendo o nó raiz (nível

Tabela 4.7 - Comparações realizadas com os grupos, sendo listado o tipo de pré-processamento feito e a quantia de casos utilizados de cada grupo para o treinamento e teste.

Filtro	Grupos*	Conjunto Treinamento	Conjunto Teste
FC	CONT (88) e COB (88)	40	48
FA	RNN (26) e RNP (48)	17	9
FA	RNN (26) e PC (61)	18	8
FA	RNP (48) e PC (61)	30	18
FA	VOL (61) e PC (61)	45	16
FA	VOL (61) e RNP (48)	30	18
FA	VOL (61) e RNN (26)	17	9
PR	NOR (54) e CHF (15)	10	5
PR	NOR (54) e APN (27)	20	7
PR	APN (27) e CHF (15)	10	5

*Entre parênteses está o total de casos de cada grupo.
FC = filtro convencional, FA = filtro adaptativo e PR = pontos removidos.
CONT - Criança com Peso Normal, COB - Criança com Sobrepeso
RNP - Recém-Nascido Prematuro, RNN - Recém-Nascido Normal
VOL - Adulto Jovem Saudável, PC - Adulto Coronariopata
NOR - Adulto com Ritmo Sinusal Normal, CHF - Adulto com Falha Congestiva Cardíaca
APN - Adulto com Insufiência Respiratória

0) da árvore de decisão. Esse nó raiz representa a medida que mais difere os dois grupos de séries temporais de intervalos RR usados. Em outras palavras, no nó raiz está a medida com maior entropia.

A Tabela 4.8 apresenta a acurácia média obtida para cada comparação de grupos e seu respectivo nó raiz. Observe que para todas as comparações com acurácia superior a 80%, são destacadas as medidas do gráfico de Poincaré e CTM como nivel 0 da árvore de decisão.

A Figura 4.9 apresenta um exemplo de árvore de decisão obtida usando o J48 para a comparação dos grupos VOL (jovens adultos saudáveis) e PC (adultos coronariopatas). Observe que para a construção da árvore a partir dessa comparação foram necessários quatro níveis de nós para classificar todo o conjunto de dados, estando no nível 0 a medida CTM para $\rho = 0,09s$. Há uma relação direta entre o número de níveis necessários para elaboração da árvore de decisão e a separabilidade dos dois conjuntos de dados. Quanto menos níveis na árvore, mais fácil separar os conjuntos, ou seja, as medidas fornecidas como entrada detectam a diferente dinâmica dos

grupos comparados.

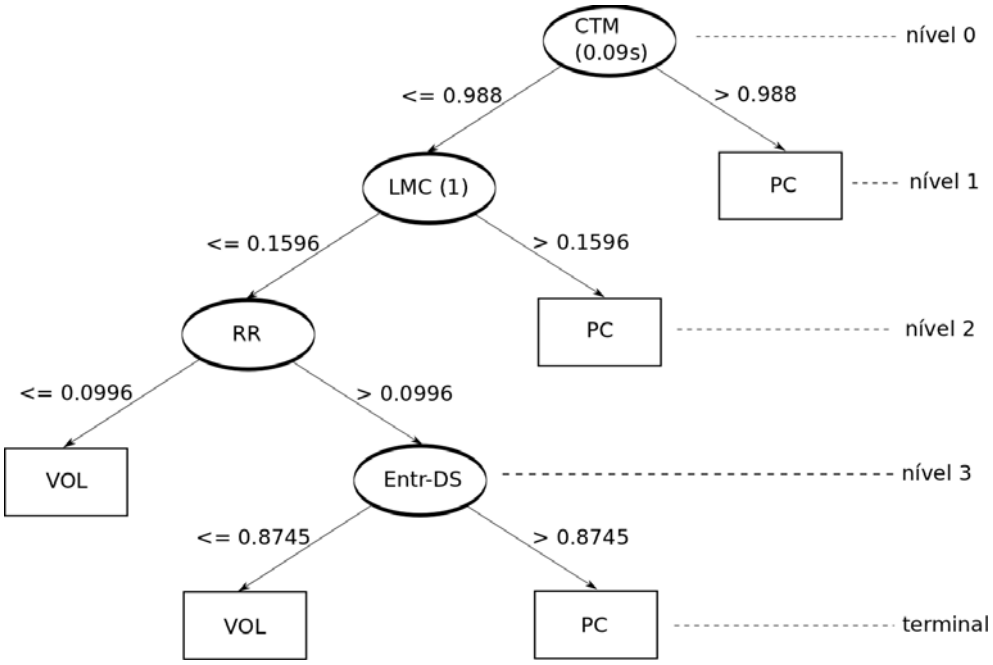

Figura 4.9 - Exemplo de árvore de decisão obtida para a comparação dos grupos VOL (jovens adultos saudáveis) e PC (adultos coronariopatas).

Tabela 4.8 - Valores de acurácia média para todas as comparações quando todas as medidas são apresentadas de uma vez como padrões de entrada. Para cada combinação há a medida que é estabelecida como nó raiz pelo classificador J48.

Comparações	Acurácia média	Nó raiz
COB e CONT	$58,75\%$	TT
RNN e RNP	$96,61\%$	SD1/SD2
RNN e PC	$95,72\%$	SD2
RNP e PC	$87,00\%$	CTM(0.01s)
VOL e PC	$82,97\%$	CTM(0.09s)
VOL e RNP	$96,13\%$	CTM(0.01s)
VOL e RNN	$94,78\%$	SD1/SD2
NOR e CHF	$88,8\%$	SD2
NOR e APN	$91,71\%$	SD1/SD2
APN e CHF	$83,00\%$	CTM(0.11s)

4.2.2 SVM

Para a análise usando o classificador SVM foram adotadas três abordagens diferentes. A primeira abordagem foi a apresentação de todas as medidas de cada comparação obtendo a acurácia média. A segunda abordagem foi apresentar apenas uma medida por vez ao classificador calculando-se a acurácia média. Essa segunda abordagem permitiu estabelecer quais foram as medidas capazes ou não de detectar diferenças entre as dinâmicas dos sistemas.

A terceira abordagem foi apresentar apenas duas medidas por vez ao classificador, estimando-se uma acurácia média ligada à duas medidas fornecidas como entrada. Foram testadas todas as combinações dois a dois de medidas para as comparações dos grupos realizadas.

Para todas as abordagens, a acurácia média foi calculada a partir de 100 execuções do algoritmo. Considerando cada comparação de grupos estudada, o algoritmo foi executado 100 vezes na primeira abordagem, 2600 vezes (ou 1700 vezes para os conjuntos Physionet e CORIA) na segunda abordagem e 32500 vezes (ou 13600 vezes para os conjuntos Physionet e CORIA) na terceira abordagem.

Primeira abordagem

Para a primeira abordagem, a Tabela 4.9 mostra os valores de acurácia média para todas as comparações de dois grupos realizadas com todas as medidas apresentadas como entrada no SVM.

Tabela 4.9 - Valores de acurácia média para todas as comparações obtidas quando todas as medidas são apresentadas juntas como entrada no SVM.

Comparações	Acurácia média
COB e CONT	$61,17\%$
RNN e RNP	$99,74\%$
RNN e PC	$97,56\%$
RNP e PC	$88,44\%$
VOL e PC	$81,75\%$
VOL e RNP	$97,28\%$
VOL e RNN	$97,44\%$
NOR e CHF	$97,5\%$
NOR e APN	$98,86\%$
APN e CHF	$95,7\%$

Observamos que apenas a comparação COB e CONT apresentou uma acurácia média abaixo de 75%. Este valor aproxima-se do obtido pelo J48.

Segunda abordagem

Na segunda abordagem, temos o valor de acurácia média para cada medida fornecida como entrada para o SVM. Para análise dos índices, convencionamos empiricamente que valores de acurácia média acima de $0,75$ (linha tracejada) indicam que os valores dos índices discriminam os grupos comparados.

A Figura 4.10 apresenta as acurácias médias para a comparação CONT (crianças de peso normal) e COB (crianças acima do peso). Observe que não há nenhuma medida capaz de detectar diferenças (superiores a $0,75$) na dinâmica desses dois conjuntos de dados.

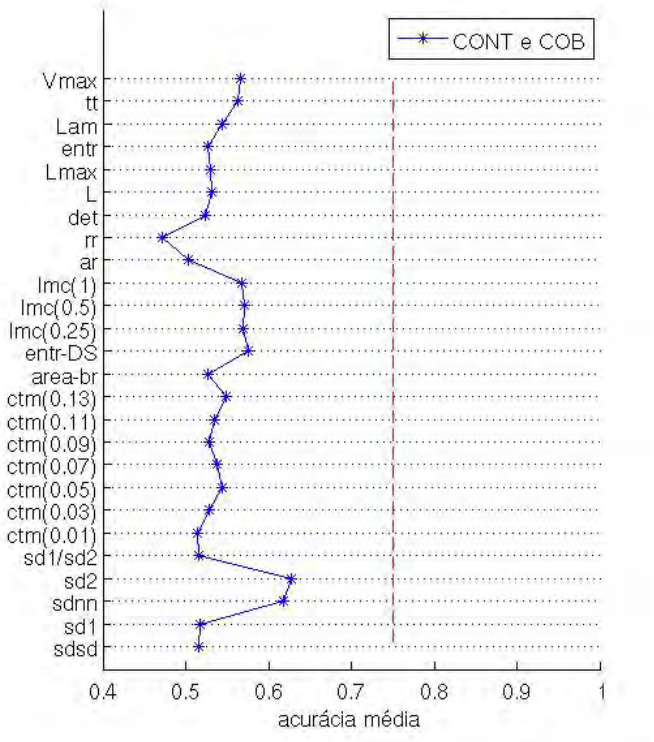

Figura 4.10 - Acurácias obtidas para cada um dos índices fornecidos como entrada no SVM para a comparação CONT e COB. **As linhas que unem os pontos do gráfico são colocadas apenas para melhor visualização dos resultados.**

Para a comparação RNP (recém-nascidos prematuros) e RNN (recém-nascidos normais) (Figura 4.11A) temos que apenas as medidas SDNN, SD2 e SD1/SD2 detectaram diferenças (superiores a $0,75$) entre os dois grupos. Sabemos que SD2 está relacionado aos intervalos RR ao longo do tempo e SD1/SD2 a razão entre as variações rápida (SD1) e de longa duração da frequência cardíaca, indicando que a diferença da dinâmica dos RNP e RNN pode estar relacionada a VFC ao longo do tempo e não na variação imediata.

Entre os grupos comparados VOL (jovens adultos saudáveis) e PC (adultos coronariopatas) (Figura 4.11B) foram observados mais medidas que detectam diferenças entre os grupos com valores de acurácia média acima de $0,75$ do que entre RNP e RNN.

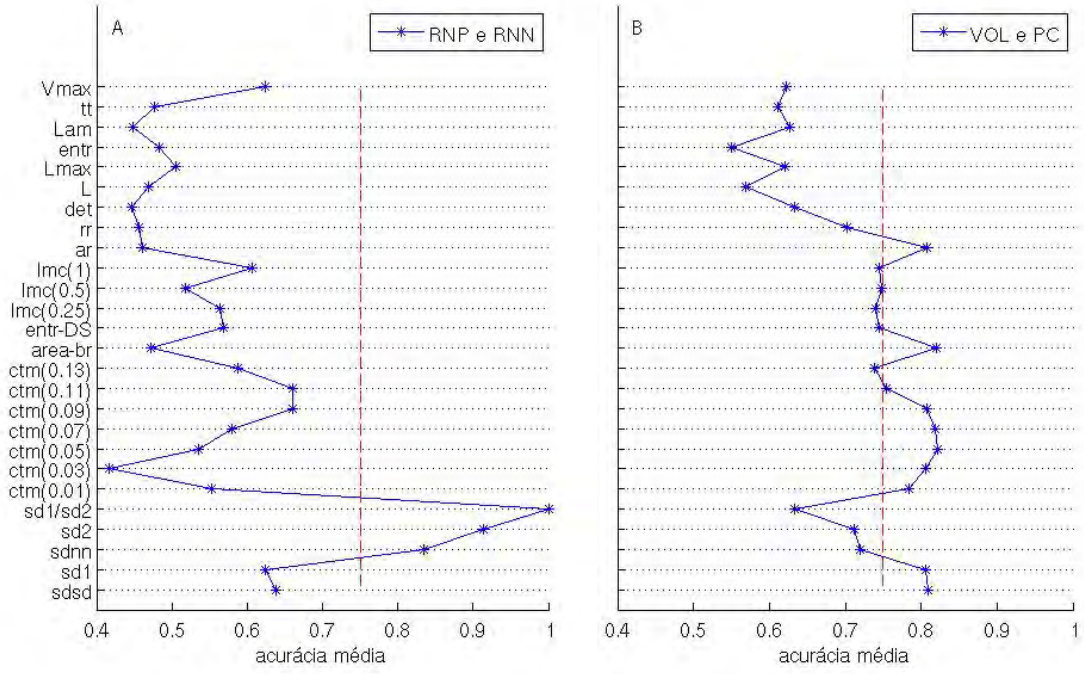

Figura 4.11 - Acurácias obtidas para cada um dos índices fornecidos como entrada no SVM para a comparação RNN e RNP (A) e para a comparação VOL e PC (B). **As linhas que unem os pontos do gráfico são colocadas apenas para melhor visualização dos resultados.**

A Figura 4.12 apresenta as acurácias obtidas pelas medidas para a comparação RNN e VOL em 4.12A e para a comparação RNN e PC em 4.12B. De uma forma geral, observe que na comparação dos grupos o conjunto de medidas utilizado detecta mais

diferenças entre RNN e VOL do que RNN e PC.

Um comportamento muito similar pode ser observado na Figura 4.13 que apresenta as acurácias obtidas pelas medidas para a comparação RNP e VOL em 4.13A e para a comparação RNP e PC em 4.13B. As medidas distinguem mais diferenças na dinâmica entre RNP e VOL do que entre RNP e PC.

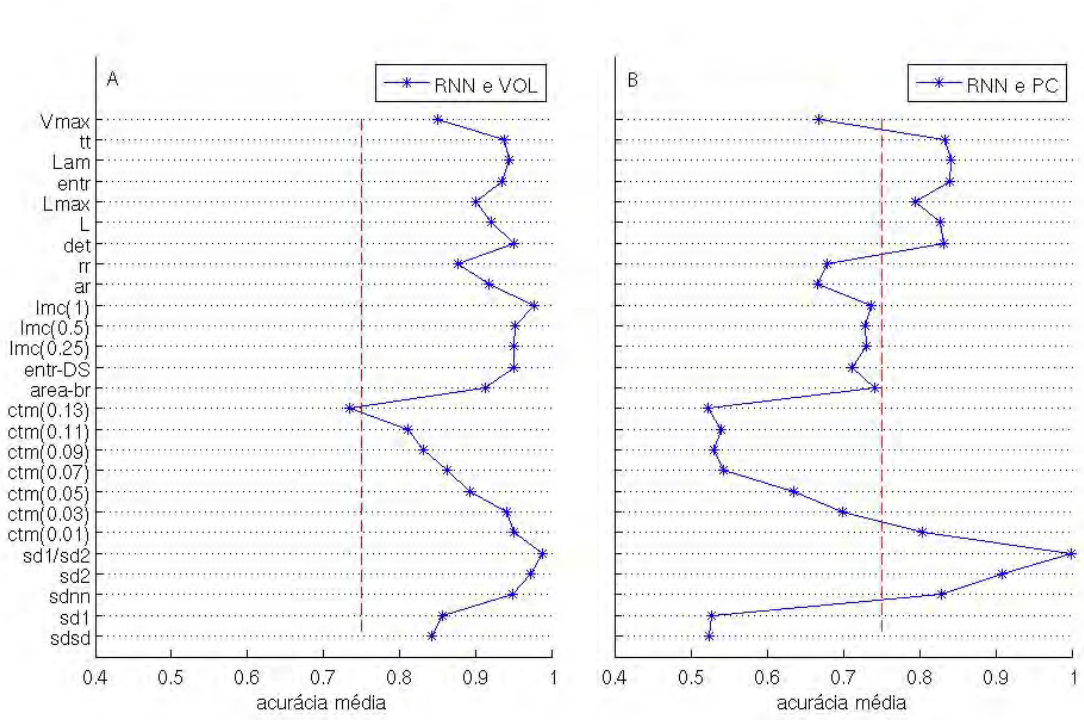

Figura 4.12 - Acurácias obtidas para cada um dos índices fornecidos como entrada no SVM para a comparação RNN e VOL (A) e para a comparação RNN e PC (B). **As linhas que unem os pontos do gráfico são colocadas apenas para melhor visualização dos resultados.**

Para os conjuntos de séries temporais de intervalos RR com pontos removidos foram calculadas apenas 17 medidas conforme já mencionado. Na Figura 4.14A temos as acurácias médias para a comparação NOR (adultos normais do Physionet) e CHF (adultos com falha cardíaca congestiva). Observe que muitas medidas apresentam acurácia média superior a $0,75$. Isso evidencia que as medidas usadas foram capazes de detectar a dinâmica diferente desses dados.

Esse mesmo desempenho não foi presente na comparação NOR e APN (adultos com insuficiência respiratória) na Figura 4.14B e na comparação CHF e APN. Ou

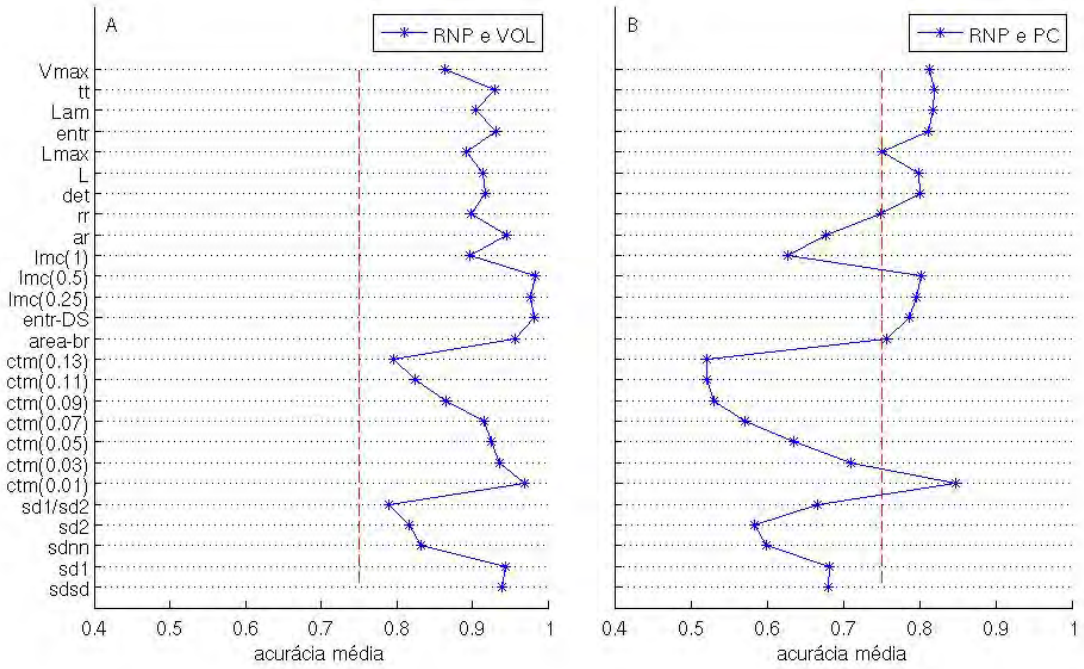

Figura 4.13 - Acurácias obtidas para cada um dos índices fornecidos como entrada no SVM para a comparação RNP e VOL (A) e para a comparação RNP e PC (B). **As linhas que unem os pontos do gráfico são colocadas apenas para melhor visualização dos resultados.**

seja, existem mais diferenças na dinâmica entre NOR e CHF do que entre NOR e APN e CHF e APN (Figura 4.15). Isso poderia ser um indíce que o diagnóstico de insuficiência respiratória para o grupo APN apresenta uma grande diversidade de VFC dos pacientes.

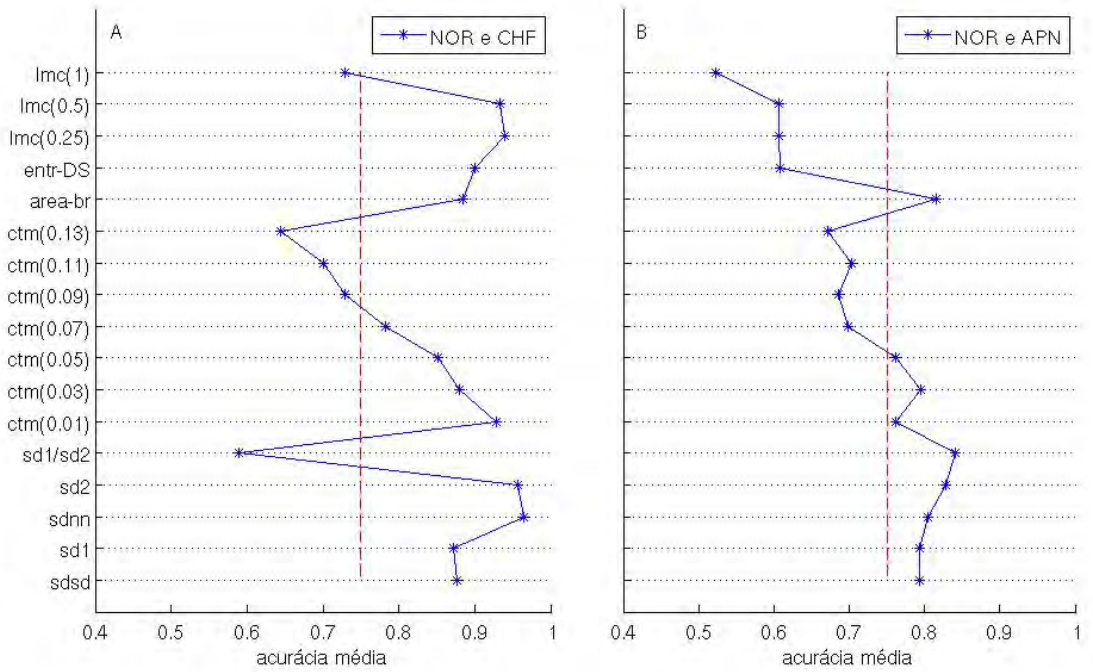

Figura 4.14 - Acurácias obtidas para cada um índices fornecidos como entrada no SVM para a comparação NOR e CHF (A) e para a comparação NOR e APN (B). **As linhas que unem os pontos do gráfico são colocadas apenas para melhor visualização dos resultados.**

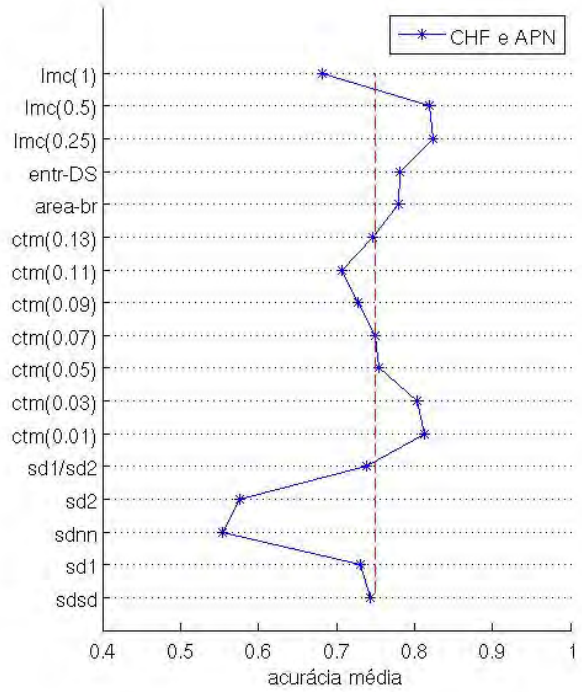

Figura 4.15 - Acurácias obtidas para cada um índices fornecidos como entrada no SVM para a comparação CHF e APN. **As linhas que unem os pontos do gráfico são colocadas apenas para melhor visualização dos resultados.**

Terceira abordagem

Para a terceira abordagem foi usada a estratégia de selecionar duas medidas como entrada no SVM. Foram usadas todas as possíveis combinações dois a dois das 26 medidas (325 combinações) extraídas dos conjuntos de tacogramas. É importante salientar que, para os conjuntos de tacogramas onde os pontos foram removidos usando o filtro adaptativo como detector, foram calculadas apenas 17 medidas, resultando em 136 diferentes combinações. Cada valor de acurácia média está associado a duas medidas fornecidas como entrada, conforme mencionado na Subseção 2.3.2.

A Figura 4.16 apresenta as acurácias médias para a comparação CONT e COB. Observe que não há nenhuma combinação de duas medidas que foram capazes de detectar diferenças (superiores a 0.75) na dinâmica desses dois conjuntos de dados. Um desempenho similar para essa comparação se apresentou também na primeira e na segunda abordagem usando SVM. Isso pode ser um indicativo de que, apesar das diferenças clínicas encontradas por um médico especialista, que os classificou em grupos distintos, a dinâmica na VFC (relacionada ao SNA) não difere significativamente entre os dois grupos

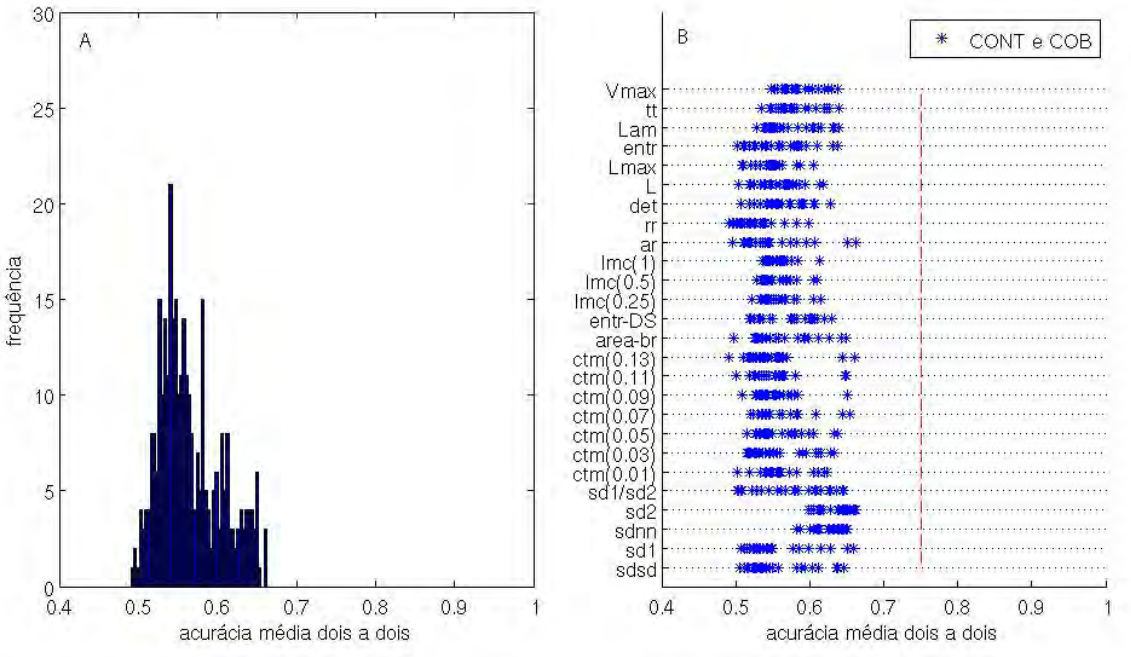

Figura 4.16 - Acurácias médias obtidas para as combinações dois a dois de 26 índices fornecidos como entrada para o SVM para CONT e COB. a) Histograma das acurácias obtidas. b) Valores das acurácias médias para combinação dois a dois de medidas.

As Figuras 4.17A e 4.17B apresentam as acurácias médias para a comparação RNP e RNN. Quando usamos combinações de duas medidas, aparecem novos valores de acurácia para medidas que na segunda abordagem não possuiam. Entretanto, as medidas que apresentam melhor desempenho (valores superiores a 0.75) continuam sendo SDNN, SD2 e SD1/SD2. Observe que nas Figuras 4.17C e 4.17D para a comparação VOL e PC as medidas que apresentam valores de acurácia inferior a 0.75 passaram a possuir acurácias médias maiores, graças as combinações de medidas realizadas.

Comparando esse resultado com o obtido na segunda abordagem, note que muitas medidas possuem valores de acurácia sensivelmente superiores. Isso ocorre, pois a combinação de duas medidas forneceu mais informações ao classificador que apenas uma como entrada.

A Figura 4.18 apresenta os valores de acurácia média para as combinações de duas medidas para as comparações RNN e VOL (A e B) e RNN e PC (C e D). Os resultados dessas combinações reforçam os resultados para essa mesma comparação com a segunda abordagem usando o SVM. Entretanto, observa-se que duas medidas como entrada no SVM fornecem mais informações resultando em valores de acurácia maiores. Um desempenho similar pode ser observado nas Figuras 4.19 para as comparações RNP e VOL e RNP e PC 4.20 para as comparações NOR e CHF e NOR e APN e 4.21 para a comparação CHF e APN.

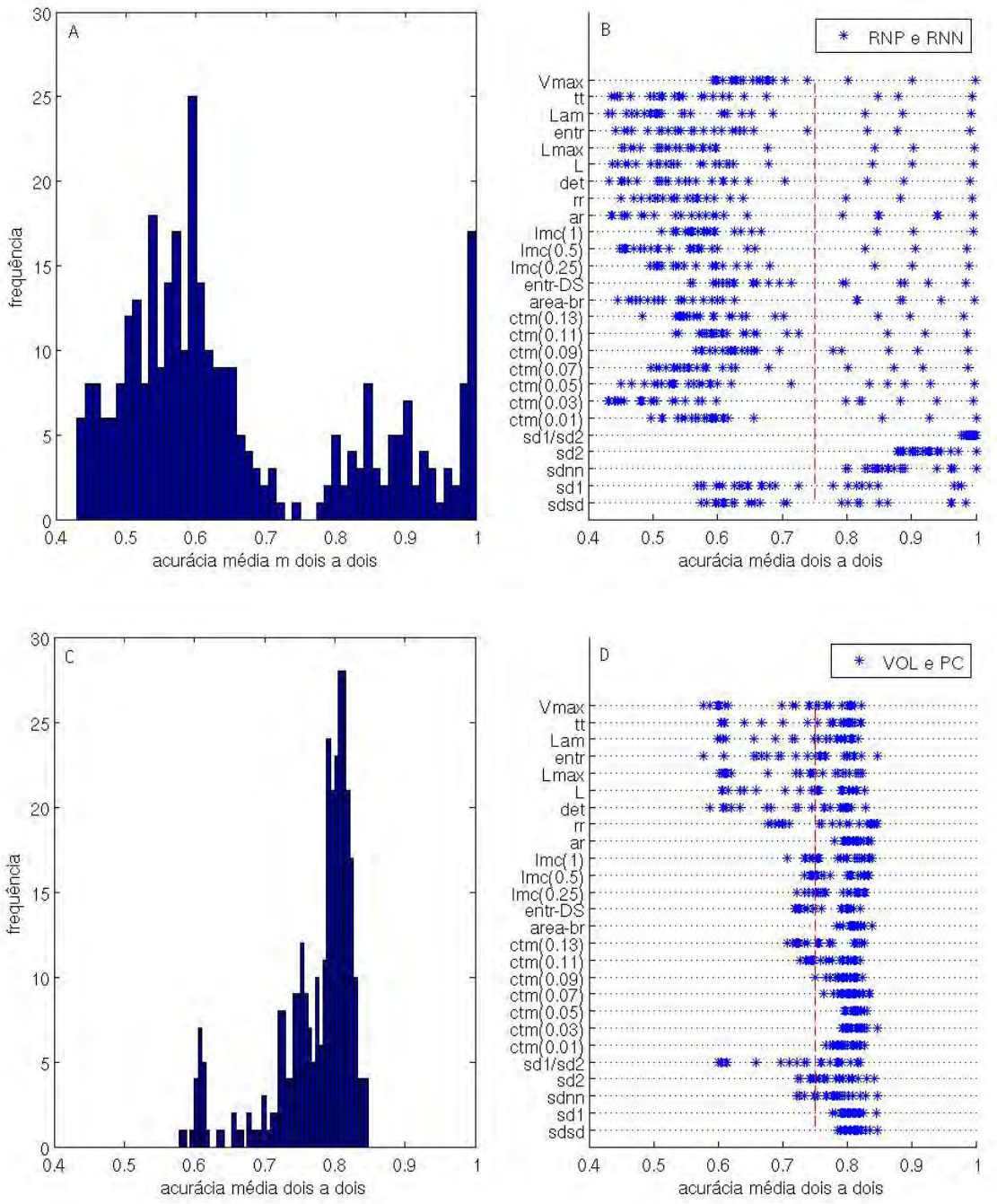

Figura 4.17 - Acurácias médias obtidas para as combinações dois a dois de 26 índices fornecidos como entrada para o SVM para RNP e RNN (A e B) e para VOL e PC (C e D). A) Histograma das acurácias obtidas para RNP e RNN. B) Valores das acurácias médias para combinação dois a dois de medidas para RNP e RNN. C) Histograma das acurácias obtidas para VOL e PC. D) Valores das acurácias médias para combinação dois a dois de medidas para VOL e PC.

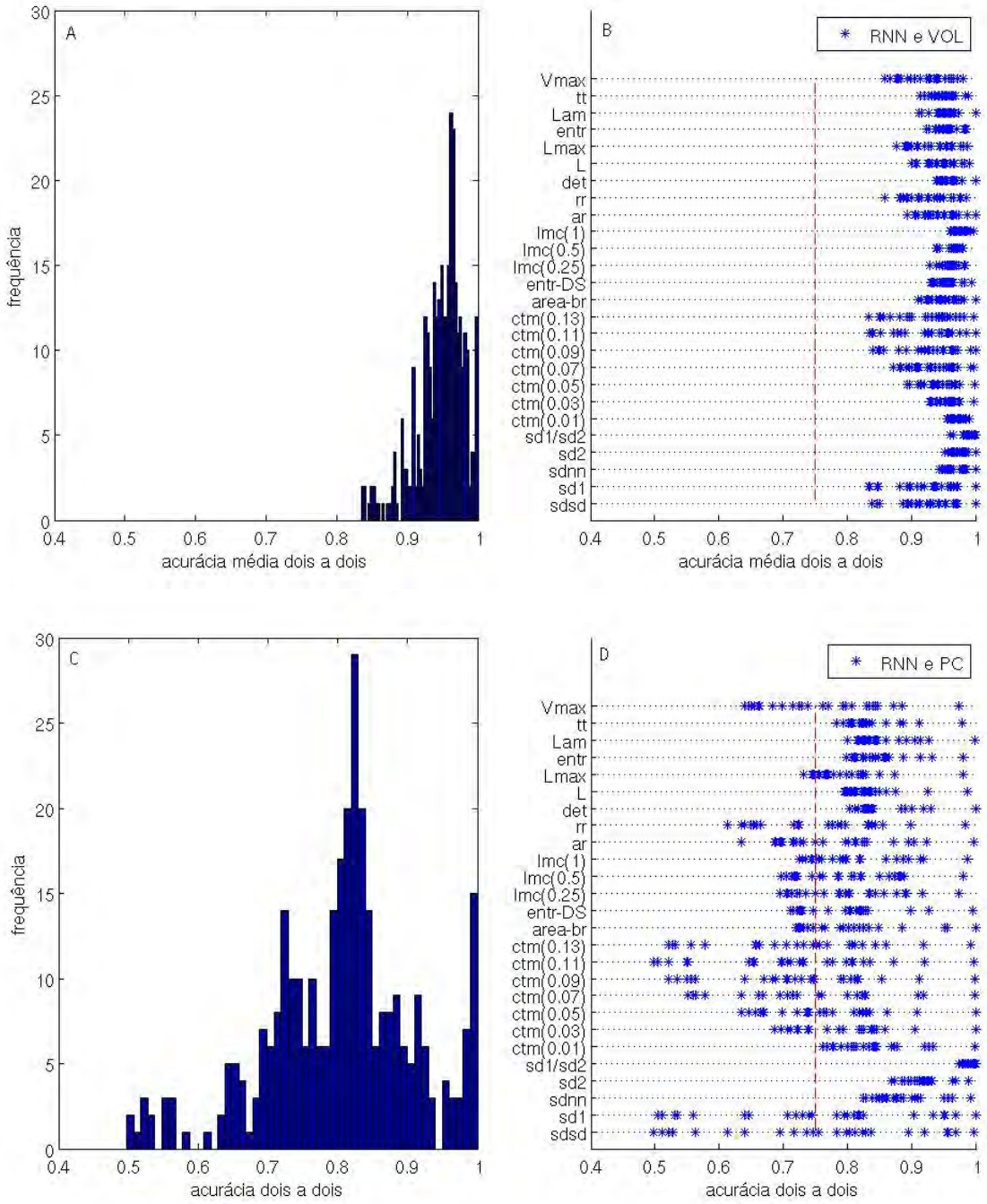

Figura 4.18 - Acurácias médias obtidas para as combinações dois a dois de 26 medidas fornecidas como entrada para o SVM para RNN e VOL (A e B) e para RNN e PC (C e D). A) Histograma das acurácias obtidas para RNN e VOL. B) Valores das acurácias médias para combinação dois a dois de medidas para RNN e VOL. C) Histograma das acurácias obtidas para RNN e PC. D) Valores das acurácias médias para combinação dois a dois de medidas para RNN e PC.

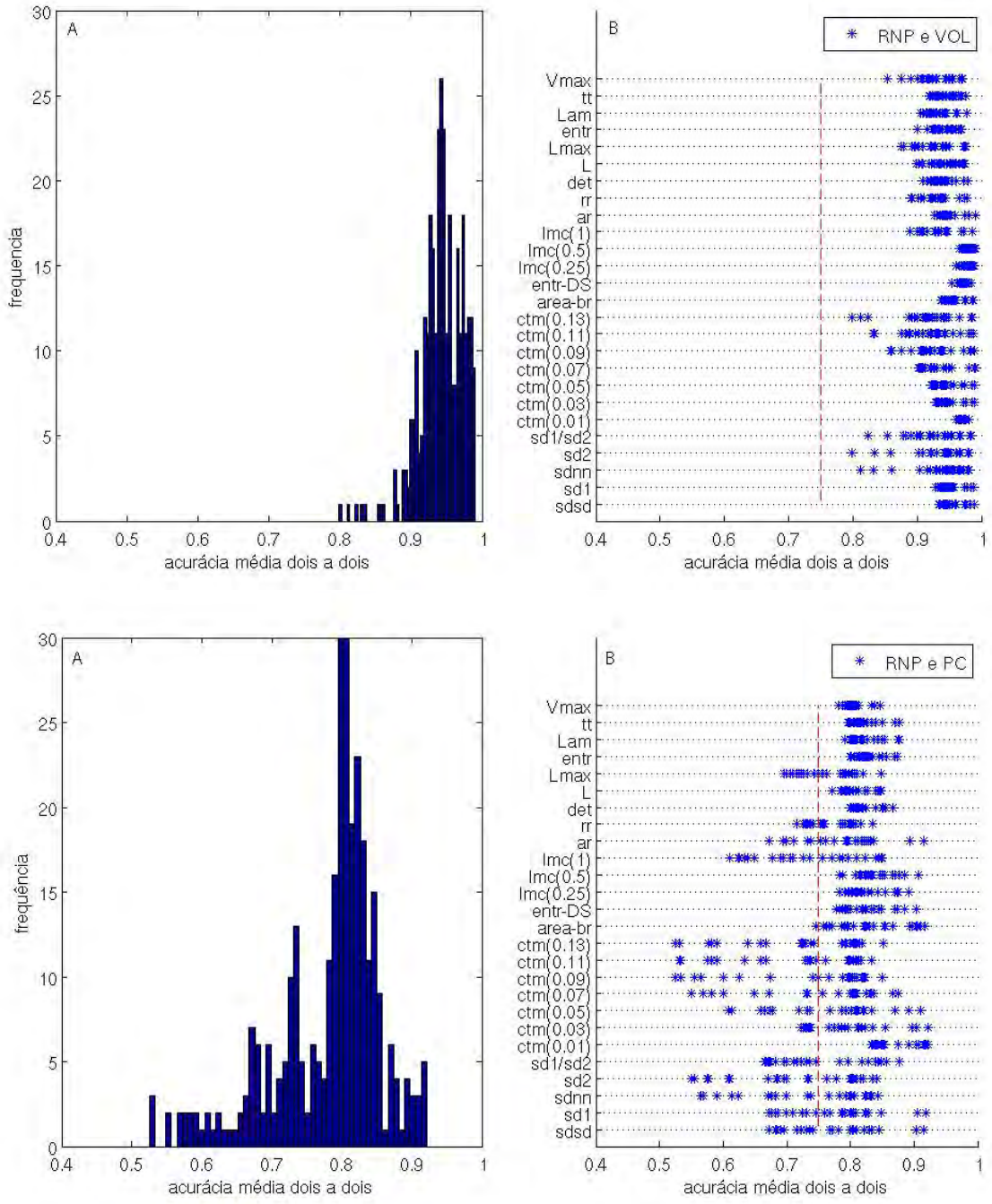

Figura 4.19 - Acurácias médias obtidas para as combinações dois a dois de 26 medidas fornecidas como entrada para o SVM para RNP e VOL (A e B) e para RNP e PC (C e D). A) Histograma das acurácias obtidas para RNP e VOL. B) Valores das acurácias médias para combinação dois a dois de medidas para RNP e VOL. C) Histograma das acurácias obtidas para RNP e PC. D) Valores das acurácias médias para combinação dois a dois de medidas para RNP e PC.

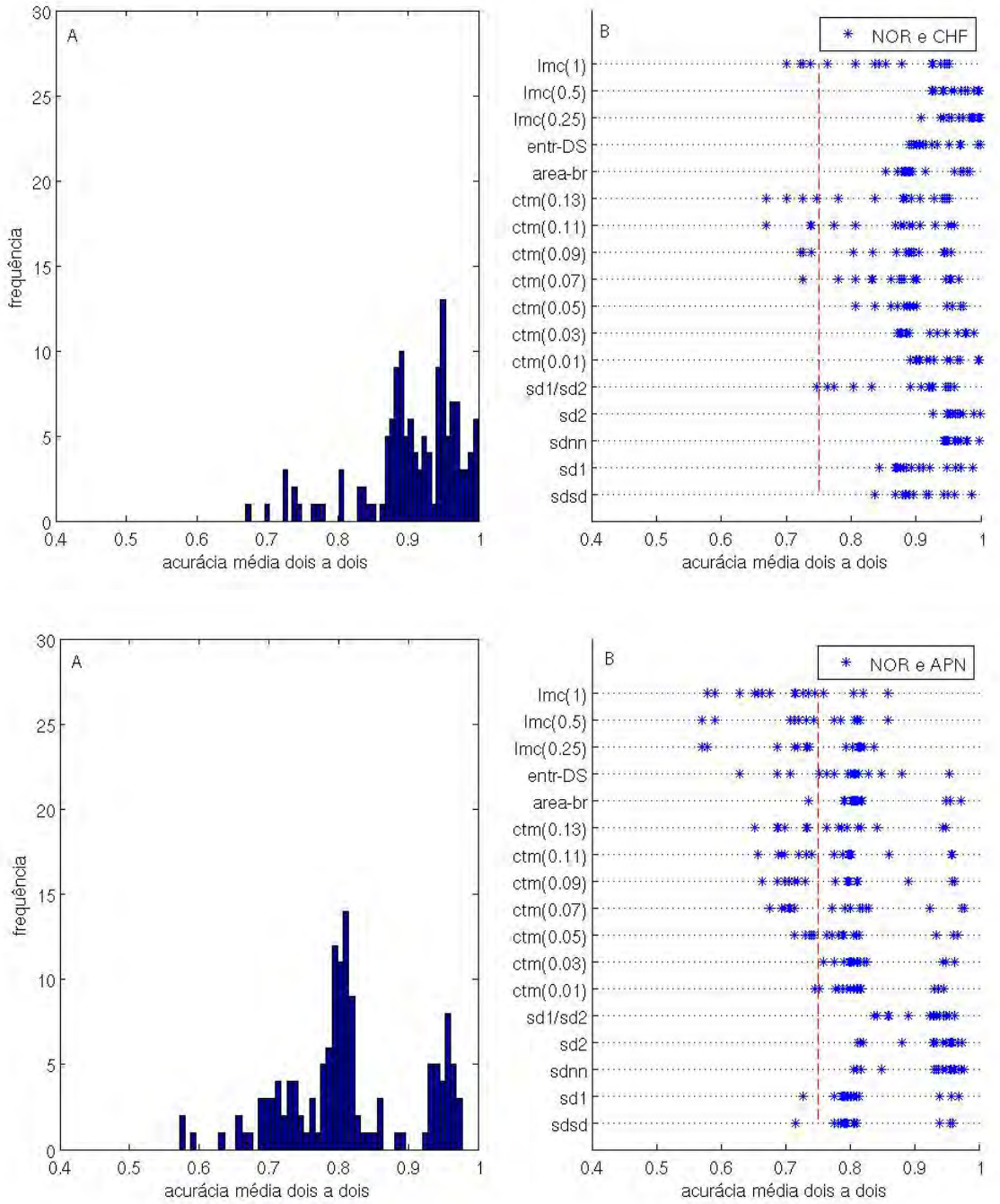

Figura 4.20 - Acurácias médias obtidas para as combinações dois a dois de 17 medidas fornecidas como entrada para o SVM para NOR e CHF (A e B) e para NOR e APN (C e D). A) Histograma das acurácias obtidas para NOR e CHF. B) Valores das acurácias médias para combinação dois a dois de medidas para NOR e CHF. C) Histograma das acurácias obtidas para NOR e APN. D) Valores das acurácias médias para combinação dois a dois de medidas para NOR e APN.

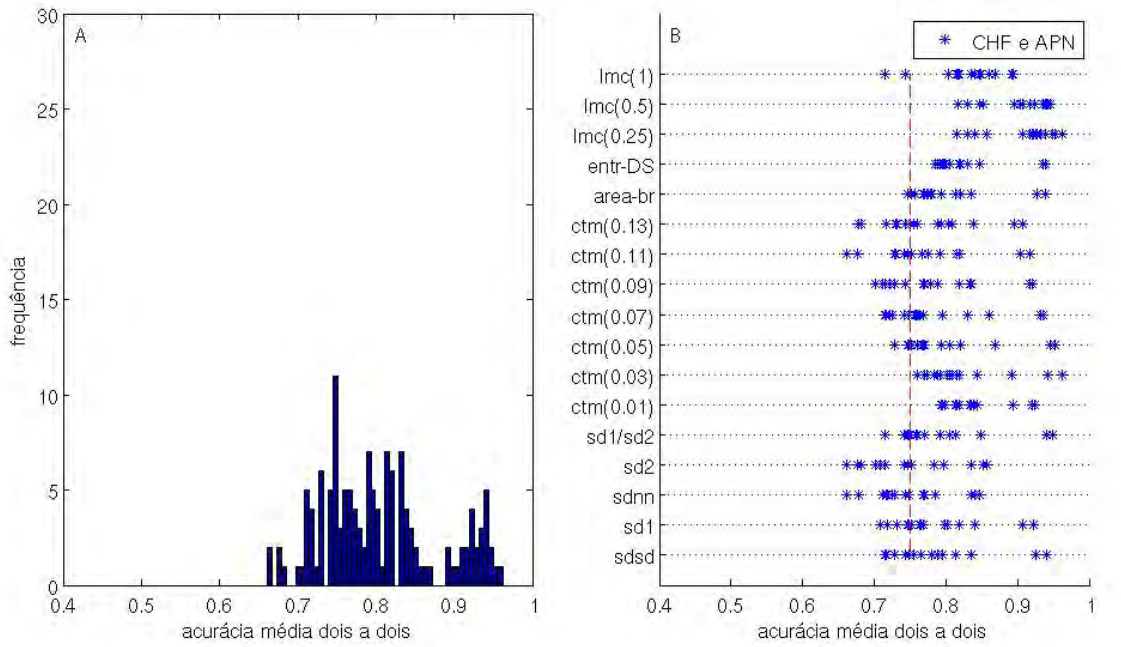

Figura 4.21 - Acurácias médias obtidas para as combinações dois a dois de 17 medidas fornecidas como entrada para o SVM para CHF e APN. A) Histograma das acurácias obtidas. B) Valores das acurácias médias para combinação dois a dois de medidas.

4.2.3 Síntese de resultados

Com a finalidade de sintetizar os resultados obtidos com o uso do classificador SVM, a Tabela 4.10 apresenta os resultados baseados nos valores de acurácia média e as principais medidas que detectaram as diferenças entre os grupos de tacogramas comparados. Na primeira coluna da Tabela estão listadas as comparações estudadas, na segunda as medidas que apresentaram acurácia média superior a $0,75$.

Tabela 4.10 - Síntese dos resultados obtidos com o classificador SVM. Para cada comparação estão listadas as medidas que apresentaram acurácia média (A) superior a $0,75$.

Comparações	**Medidas com** $A > 0,75$
COB e CONT	**não há**
RNN e RNP	SDNN, SD2 e SD1/SD2
VOL e PC	**exceção para** SD2 e medidas RQA
RNN e VOL	**todas**
RNN e PC	SDNN, SD2, SD1/SD2, CTM(00.01), DET, L, ENTR, Lam, TT
RNP e VOL	**todas**
RNP e PC	CTM(0.01), entr-DS, LMC(0.25), DET, L, ENTR, Lam, TT, Vmax
NOR e CHF	**exceção para** CTM(0.07), CTM(0.11), CTM(0.13), LMC(1)
NOR e APN	SDNN, SD2, SD1/SD2, CTM(0.03)
APN e CHF	CTM(0.01), CTM(0.03), entr-DS, LMC(0.25), LMC(0.5)

Observe que as medidas do gráfico de Poincaré aparecem com destaque em três comparações estudadas: RNN e RNP; RNN e PC; NOR e APN. Evidenciando a importância dessas medidas em diferenciar a variabilidade de curta e longa duração nos grupos de tacogramas estudados. Em especial, sabemos que a medida SD2 do gráfico de Poincaré está relacionada, que parece na distinção das comparações destacadas, está relacionada à variabilidade de longa duração. Isso nos fornece indícios que entre essas comparações há diferenças do sistema nervoso autônomo que evoluem conforme a faixa etária dos indivíduos.

Por fim, detectamos que não houve medidas com acurácia média superior a $0,75$ para a comparação CONT e COB. Isso pode indicar que os critérios clínicos diagnosticados pelo médico (que os classificou em dois grupos distintos) ainda não provocaram diferenças na dinâmica das séries temporais. Outra hipótese é que essas medidas não são sensíveis o suficiente para detectarem essas diferenças entre esses dois sistemas semelhantes.

4.3 Considerações sobre o Capítulo

Nesse capítulo foram apresentados os resultados relacionados à análise das séries temporais de intervalos RR usando a metodologia proposta. Os resultados foram divididos conforme o tipo de análise realizada: análise dos dados com a utilização do filtro adaptativo e a caracterização do conjunto de tacogramas.

Em relação à primeira parte dos resultados, a análise dos dados mostrou que o uso do filtro adaptativo é estatisticamente equivalente ao uso do filtro convencional pelo especialista. Portanto, recomendamos seu uso para o pré-processamento dos dados auxiliando a análise posterior. Entretanto, é necessário que o ajuste dos parâmetros do filtro seja realizado adequadamente. Sugerimos que os parâmetros sejam ajustados tendo como referência o padrão-ouro (filtragem convencional realizada por um especialista). Nesse trabalho, a escolha dos parâmetros deu-se empiricamente tendo como referência as séries filtradas pelo especialista.

Nosso objetivo foi usar o filtro para pré-processar as séries de intervalos RR desconsiderando artefatos e arritmias para que a análise ocorresse apenas nos tacogramas com ritmo sinusal. Esse trabalho inédito já foi aceito para publicação no periódico *Medical Engineering & Physics* (`http://ees.elsevier.com/mep/`).

A segunda parte dos resultados, a caracterização dos conjuntos de séries temporais usando os classificadores J48 e SVM, mostrou que não há uma única medida (daquelas utilizadas nesse trabalho) que seja sempre capaz de detectar as diferenças na dinâmica dos grupos analisados. Mas que, índices diferentes podem detectar diferenças nas comparações de grupos realizada, diferentes medidas podem detectar a dinâmica do conjunto de dados e em outra comparação não apresentarem o mesmo desempenho.

Nesse contexto, verificamos também que podem existir grupos diagnosticados com diferente classificação clínica e essa distinção não estar muito evidente na análise da VFC, como por exemplo, no grupos CONT e COB, nos quais não foram detectadas

diferenças significativas em suas dinâmicas. Por outro lado, podem existir grupos que apresentam uma dinâmica muito diferente, como por exemplo, as comparações dos grupos RNP e VOL e dos grupos RNN e VOL nas quais todas as medidas apresentaram acurácias superiores a 0.75.

As diferenças existentes nos diferentes grupos de tacogramas, sugerem uma relação entre *variabilidade* × *faixa etária* dos indivíduos. A Figura 4.22 ilustra essa possível relação, sendo que os eixos não apresentam uma escala fixa, apenas ilustram que o tempo aumenta da esquerda para direita e a variabilidade de baixo para cima. Nessa Figura os grupos de tacogramas estudados foram posicionados no gráfico, mostrando que existem diferenças na faixa etária de cada grupo de indivíduos. Salientamos que essas diferenças não foram precisas.

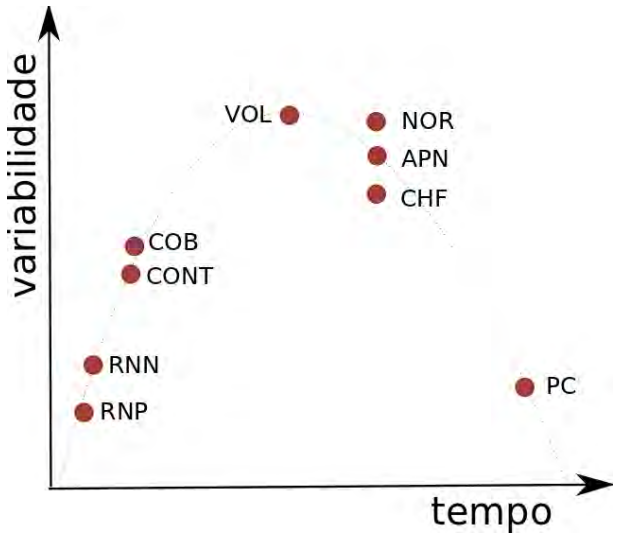

Figura 4.22 - Ilustração da relação entre variabilidade × tempo. Os eixos não possuem uma escala precisa, são apenas indicações que de o tempo (faixa etária) aumenta da esquerda para direita e que a variabilidade aumenta de baixo para cima.

Analisando os resultados obtidos na Seção 4.2, observamos que os grupos RNP e RNN apresentaram algumas medidas capazes de diferenciá-los e que ambos apresentaram mais diferenças entre eles e o grupo VOL (adultos jovens saudáveis) do que ente eles e o grupo PC (adultos coronariopatas). Sendo que RNN e PC foram ligeiramente mais distintos do que RNP e PC. Isso mostra que, apesar da faixa etária ser diferente, a variabilidade entre esses grupos não foi tão distinta.

Já as diferenças entre CONT e COB não foram suficientemente grandes para dizer

que esses dois grupos possuem dinâmicas distintas. O mesmo ocorreu para NOR (adultos normais) e APN (pacientes com insuficiência respiratória) e APN e CHF (pacientes com falha cardíaca congestiva) e APN. As comparações VOL e PC e NOR e CHF evidenciaram que a variabilidade é distinta entre os grupos.

Por fim, para que uma análise em relação à evolução da VFC seja estimada, uma vez que, possuímos conjuntos de dados de diferentes faixas etárias e condições clínicas, serão necessários mais avaliações e posteriores comparações de grupos a serem realizadas.

5 SÉRIES TEMPORAIS DE VENTO ZONAL

A metodologia elaborada caracteriza os sistemas com dinâmicas semelhantes, por isso, testamos a metodologia em mais um grupo de sinais. Neste Capítulo é apresentada a aplicação da metodologia proposta em um segundo estudo de caso: discriminação entre dois grupos de séries temporais de vento zonal (oriundas de radares meteóricos). Na Seção 5.1 é apresentada a definição de séries temporais de vento zonal e como são captadas, a Seção 5.2 apresenta o banco de dados do INPE utilizado, a Seção 5.3 aborda o pré-processamento do conjunto de dados. Por fim, a Seção 5.4 apresenta os resultados obtidos com a caracterização usando o J48 e o SVM.

5.1 Definição

As séries temporais de vento são oriundas de pesquisas do INPE usando radar meteórico. Esse radar detecta a trilha de meteoros quando eles entram na atmosfera terrestre. Podem ser detectados: fluxo de meteoros, o vento neutro e o coeficiente de difusão ambipolar, entre 70 e 110 km de altura na região mesosférica (WRASSE, 2004).

Quando um metorito penetra na atmosfera, o radar pode detectar o traço de gás ionizado deixado por ele, após sua rápida evaporação (WRASSE, 2004). Esse traço reflete um curto pulso de energia em ondas de rádio que pode ser rastreado por um conjunto de antenas receptoras. A Figura 5.1 ilustra esse processo de detecção.

Neste trabalho usamos séries temporais de vento obtidas por dois radares meteóricos: um em Cachoeira Paulista (23^{o}S, 45^{o}O) (WRASSE, 2004) e outro na Base Brasileira na Antártica, Comandante Ferraz (62.1^{o}S, 58.7^{o}O), localizada na Ilha do Rei George (FRITTS et al., 2012).

O radar meteórico instalado no Campus do INPE em Cachoeira Paulista, detecta os traços de meteoros sobre todo o céu, e opera automaticamente 24 horas por dia, detectando cerca de 3000 a 6000 meteoros observávei por dia. O radar meteórico instalado na Antártica em 2010 realiza medições de ventos médios, marés e fluxo de momentum devidos às ondas de gravidade. Esse radar detecta aproximadamente 8500 meteoros por dia (FRITTS et al., 2012).

As medidas de vento obtidas com o radar meteórico são denominadas de vento neutro. Há duas componentes no vento neutro: zonal e meridional que são determinadas pela análise dos sinais refletidos pelos traços meteóricos. Na estimativa desse vento

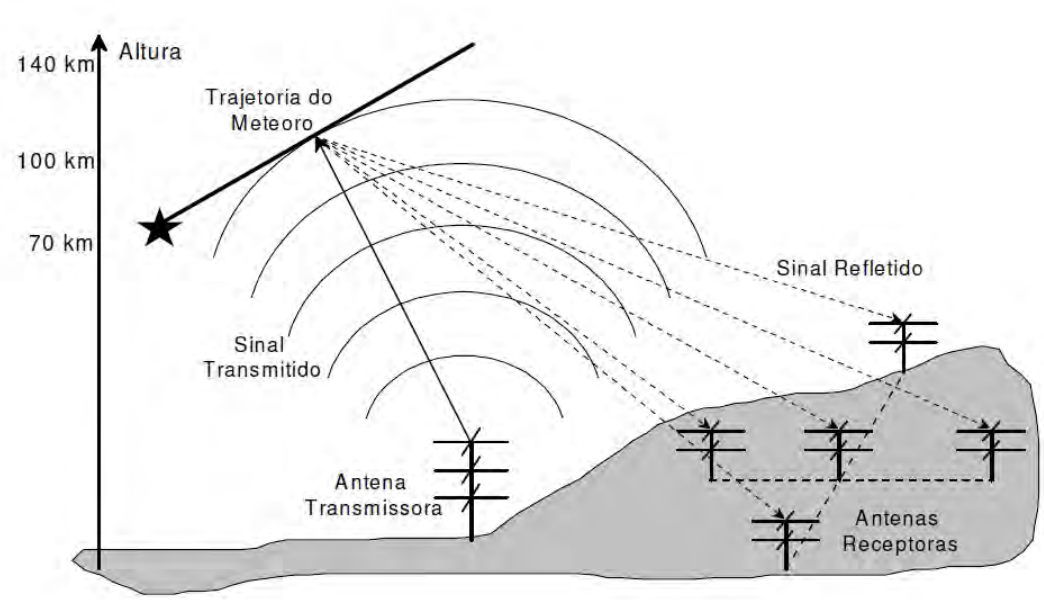

Figura 5.1 - Operação do radar meteórico. Extraído de Wrasse (2004).

torna-se necessário conhecer algumas características dos traços meteóricos, como distância entre o traço meteórico detectado e o radar, o ângulo de entrada e a direção dos traços (WRASSE, 2004).

Para medir a distância entre o traço meteórico e o observatório é necessário determinar o intervalo de tempo entre o sinal transmitido pelo radar e o sinal refletido. Considerando que o índice de refração do meio seja igual a um, podemos descrever a distância do traço meteórico em relação à base como:

$$R = \mathrm{c}T_r/2, \tag{5.1}$$

sendo c a velocidade da luz no vácuo, T_r é o intervalo de tempo entre o pulso transmitido pelo radar e o eco ser detectado. O número 2 no denominador se deve ao fato de que o tempo T_r é o tempo de ida e volta da onda, correspondente à distância R entre o radar e a trilha do meteoro. Uma breve revisão da técnica de medida de vento neutro a partir do radar meteórico é fornecida a seguir, tendo como base o trabalho de Wrasse (2004).

Depois de estabelecer a distância entre o traço meteórico e o radar, calcula-se a altura:

$$z = R\,\cos\theta \tag{5.2}$$

sendo θ é o ângulo medido a partir do zênite[1]. Estimando, por exemplo, que os meteros ocorram entre 78 a 110 km de altura, considerando uma altura máxima de 130 km, o maior ângulo será $\theta \approx 53°$.

A determinação do vento neutro na média atmosfera ocorre pelo deslocamento Doppler[2] do sinal refletido pelo rastro do meteoro. Os meteoróides produzem traços ionizados, que permanecem um certo tempo na atmosfera, possibilitando a estimativa do vento neutro nesta região (tempo tipicamente $> 0,1s$). Se o resultado obtido do deslocamento da frequência Doppler for positivo, significa aproximação em relação ao observador e negativo, o afastamento (WRASSE, 2004).

Para calcular o ângulo de entrada do meteoro é usado um interferômetro[3] composto de cinco antenas receptoras (ver Figura 5.2).

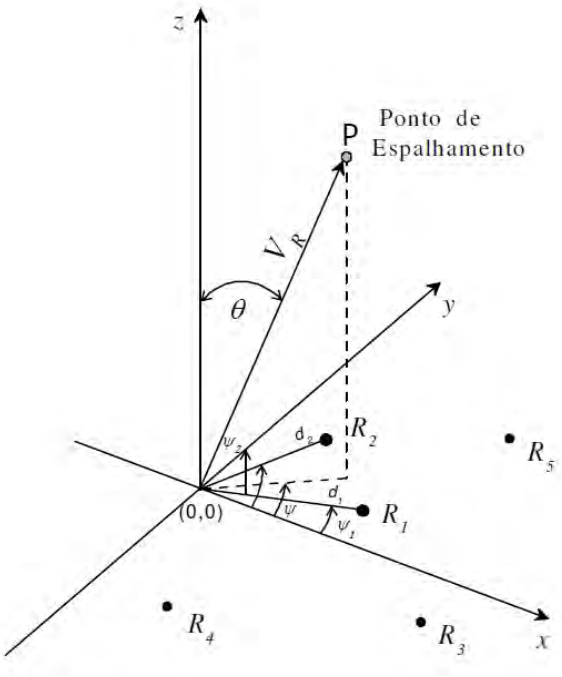

Figura 5.2 - Ilustração do sistema de detecção do ângulo de entrada do meteoro. Extraído de Wrasse (2004).

[1]O termo zênite é um dos três pontos referenciais (ponto mais alto) que formam um triângulo de posição de onde se encontra diretamente o observador (FARIA, 2007).

[2]Efeito Doppler é o efeito de modificação na frequência percebida pelo receptor de uma onda emitida por uma fonte, devido ao movimento relativo entre a fonte e o receptor (FARIA, 2007).

[3]Interferômetro é um aparelho que determina as medidas de ângulos e distâncias usando a interferência das ondas eletromagnéticas no instante que elas interagem entre si (KNIGHT, 2009).

Na Figura 5.2 o eixo x é a coordenada na direção leste, o y na direção norte e z a altura da direção vertical. Um conjunto de possíveis direções é fornecido pelas antenas R1, R2, R3 e R4 onde R5 otimiza a ambigüidade do ângulo de entrada. O ponto P é o ponto de reflexão especular[4] na direção (θ, ψ), representando os ângulos zenital e azimutal respectivamente, em relação à origem.

Após estabelecer a distância, os ângulos zenital e azimutal, tem-se um sistema de coordenadas tridimensional, que após uma série de cálculos[5] obtem-se as componentes do vento neutro nas direções zonal, meridional e vertical, permitindo determinar as variações no vento neutro em função da altura e do tempo.

5.2 Banco de dados

O estudo da velocidade (m/s) do vento neutro permite compreender a dinâmica da alta atmosfera. É usado para estudos de maré atmosférica[6] e ondas planetárias[7] (em grande escala espacial e temporal) e também para determinar as condições de propagação vertical de ondas de gravidade[8] (em pequena escala espacial e curto período) (ANDRIOLI, 2012).

Para esse estudo, usamos a componente zonal do vento neutro medido na cidade de Cachoeira Paulista no Estado de São Paulo e na Estação Comandante Ferraz na Antártica, caracterizando assim, dois grupos distintos de séries temporais. A escolha da componente zonal deve-se ao fato que essa componente melhor definida do que a componente meridional. Quando o valor medido (velocidade em m/s) for negativo, significa que a direção do vento está para oeste e quando for positivo a direção do vento está para leste.

Os dois conjuntos de séries temporais da componente zonal foram obtidas para uma altitude de aproximadamente 90 km (máximo de detecção de meteoros). A escolha dessa altura está relacionada à menor probabilidade de erros, uma vez que há maior probabilidade da ocorrência de meteoros.

[4]Reflexão especular é o reflexo espelhado da luz em uma superfície, onde uma única direção de entrada reflete em uma única saída (KNIGHT, 2009).

[5]Para melhor detalhamento sobre a determinação do vento neutro consultar Wrasse (2004) e Fritts et al. (2012).

[6]Marés atmosféricas são ondas atmosféricas que possuem escala espacial global, porém os períodos são sub-harmônicos do dia solar ou lunar (24h, 12h, 8h) (BAGESTON, 2010).

[7]Ondas planetárias são ondas de escala planetária com grandes estruturas horizontais, da ordem do diâmetro terrestre, e períodos maiores que um dia (BAGESTON, 2010).

[8]Ondas de gravidade são ondas com períodos que variam de alguns minutos a horas, onde o limite inferior dessas ondas é o período de Brünt Väisälä, que na mesosfera superior é tipicamente da ordem de cinco minutos (BAGESTON, 2010).

As medidas de Cachoeira Paulista foram coletadas durante os meses de abril e maio de 2006 e abril e maio de 2008 em uma altura de 90 km. No total, o conjunto de séries temporais de vento zonal é composto por 32 séries. Cada série, em média, é formada pela junção de quatro dias de medidas, sendo que para cada dia, em geral, obteve-se 24 medidas.

As medidas da Estação Comandante Ferraz foram coletadas durante os meses de abril e maio de 2010 e de 2011 na altura de 91 km. No total, foram obtidas 15 séries temporais, compostas de quatro dias cada e contendo em média 24 medidas por dia.

5.3 Pré-processamento das séries

Para este segundo estudo de caso, análise das séries temporais de vento zonal, o pré-processamento utilizado é a interpolação dos dados. A cada dia são fornecidas pelos radares meteóricos, em média, 24 medidas das componentes do vento neutro (zonal e meridional). Entretanto, podem ocorrer alguns problemas de detecção ou possíveis falhas no equipamento impedindo essa frequência de medições (ao longo do dia ou dos dias consecutivos).

Outro fator associado a esse conjunto de dados é que para determinadas ferramentas utilizadas na análise (como por exemplo, as medidas de RQA) há uma limitação em relação ao tamanho mínimo da série temporal, sendo necessário uma quantia superior a 100 pontos (análise feita empiricamente).

Em geral, cada série temporal analisada nesse conjunto contém 96 pontos, pois são formadas a partir de medições fornecidas em cada quatro dias sucessivos. Entretanto, devido aos fatores mencionados anteriormente, usamos a interpolação com a finalidade de reamostrar os dados para as ferramentas da metodologia proposta.

O tipo de interpolação adotada é a *spline*. O termo *spline* vem de uma régua elástica que pode ser curvada passando por um determinado conjunto de dados. De acordo com a teoria da elasticidade essa curva (gerada pela régua) é aproximada como uma função por partes, cada qual com um polinômio cúbico (RUGGIERO; LOPES, 2004).

Considere o exemplo de spline apresentado na Figura 5.3. A curva é delimitada pelos nós a e d. Variando as posições dos nós b e c, podemos alterar a inclinação da curva, não interferindo nos delimitadores. Esse conjunto de nós pode ser denominado de pontos de controle. Quando a curva passa por todos os pontos de controle chama-se de spline de interpolação e quando passa perto de todos os pontos de controle denomina-se spline de aproximação (RUGGIERO; LOPES, 2004).

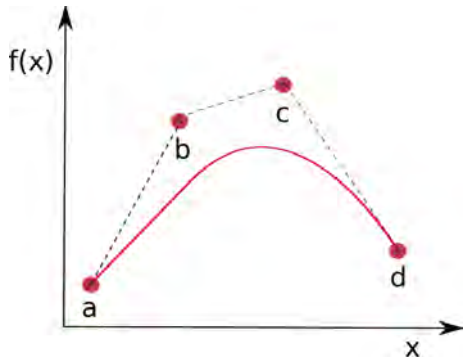

Figura 5.3 - Exemplo de spline evidenciando seus pontos de controle.

Neste trabalho é usado a spline cúbica (de interpolação) disponibilizada no Matlab®. A Figura 5.4 apresenta um exemplo de série temporal de vento zonal com 72 pontos, composta de três dias de medidas sucessivas (observe que para cada dia há 24 medidas). A mesma série é reamostrada com mais pontos, permitindo assim, a análise correta pelos métodos propostos.

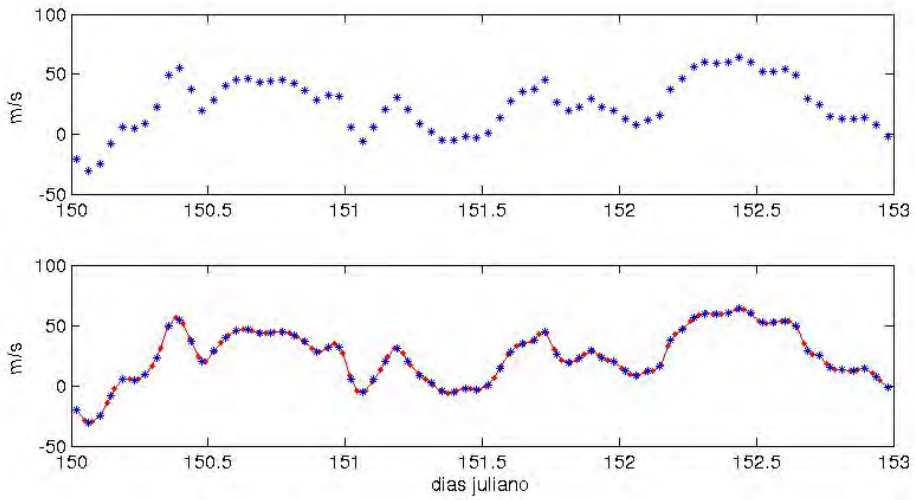

Figura 5.4 - Série temporal de vento zonal com 72 medidas obtida do radar meteórico de Cachoeira Paulista no mês de Maio de 2008. Na parte superior: série original não interpolada. Na parte inferior: série interpolada com spline com 108 pontos.

5.4 Resultados

Nessa seção são apresentados os resultados obtidos com o conjunto de séries temporais de vento zonal do INPE a partir da coleta de dados oriundos dos radares meteóricos em Cachoeira Paulista (CP) e da Estação na Antártica Comandante Ferraz (CF). Aqui são apresentados os resultados obtidos da caracterização das medidas extraídas das séries temporais usando as técnicas de mineração de dados.

O conjunto de dados é composto de 32 séries temporais de vento zonal de CP e 15 séries temporais de vento zonal de CF. O conjunto de treinamento para os classificadores de mineração de dados (J48 e SVM) é formado por 9 séries temporais de cada localidade, enquanto o conjunto de teste por 6 séries temporais. Para combinação de medidas apresentada como entrada nos classificadores é calculada a acurácia média. A acurácia média é obtida após 100 iterações, sendo que os casos que compõem os conjuntos de treinamento e teste de dados são escolhidos aleatoriamente a cada nova iteração. Nesse segundo estudo de caso são extraídas 23 medidas listadas na Tabela 5.1.

Na Subseção 5.4.1 são apresentados os resultados obtidos com o classificador J48 e na Subseção 5.4.2 com o SVM.

Tabela 5.1 - Medidas de sistemas dinâmicos usadas como entrada nas técnicas de IA.

N^o	Medida	N^o	Medida
1	sdsd	2	sd1
3	sdnn	4	sd2
5	sd1/sd2	6	ctm(2)
7	ctm(4)	8	ctm(6)
9	ctm(8)	10	ctm(10)
11	ctm(12)	12	ctm(14)
13	ctm(16)	14	ctm(18)
15	ctm(20)	16	entr-DS
17	rr	18	det
19	L	20	Lmax
21	entr	22	Lam
23	tt		

5.4.1 J48

Na análise com o J48, todas as 23 medidas são apresentadas juntas como padrões de entrada. Primeiramente para os casos do conjunto de treinamento, após para os

casos de testes. A acurácia média é calculada a partir da média das 100 execuções do classificador.

O valor da acurácia médio obtido é $0,8508$, sendo que a medida escolhida pelo classificador como o nó raiz é LAM.

5.4.2 SVM

Nessa Subseção são apresentados os resultados obtidos com a metodologia proposta usando apenas o classificador SVM. Os resultados são divididos conforme a abordagem adotada para apresentação das medidas obtidas a partir das séries temporais de vento zonal.

Primeira abordagem

Na primeira abordagem todas as 23 medidas são apresentadas ao SVM como padrões de entrada ao mesmo tempo. A acurácia média obtida foi de $0,9150$.

Esse valor de acurácia caracteriza a capacidade das medidas de distinguirem os dois grupos de séries temporais de vento zonal utilizadas. Em outras palavras, há diferenças entre os conjuntos de séries que esse conjunto de medidas é capaz de detectar.

Segunda abordagem

Na segunda abordagem temos a acurácia média para cada uma das 23 medidas usadas. A Figura 5.5 apresenta os valores de acurácia média para cada uma das medidas.

Observe na Figura 5.5 que as medidas separadamente não apresentam valores de acurácia acima de $0,9$. Do conjunto de 23 medidas, apenas 12 apresentam acurácia $\geqslant 0,75$ e 6 medidas apresentam acurácia $\geqslant 0,8$. Isso mostra que, considerando esse conjunto de séries temporais, poucas medidas individualmente conseguem distinguir as diferentes dinâmicas presentes nos dados.

Terceira abordagem

Na terceira abordagem temos todas as combinações de duas medidas apresentadas como padrões de entrada para o classificador. Ao todo, temos 253 possíveis combinações de duas medidas (do total de 23). O valor de acurácia média (de 100 execuções para cada combinação) é atribuído para as duas medidas dadas ao classificador.

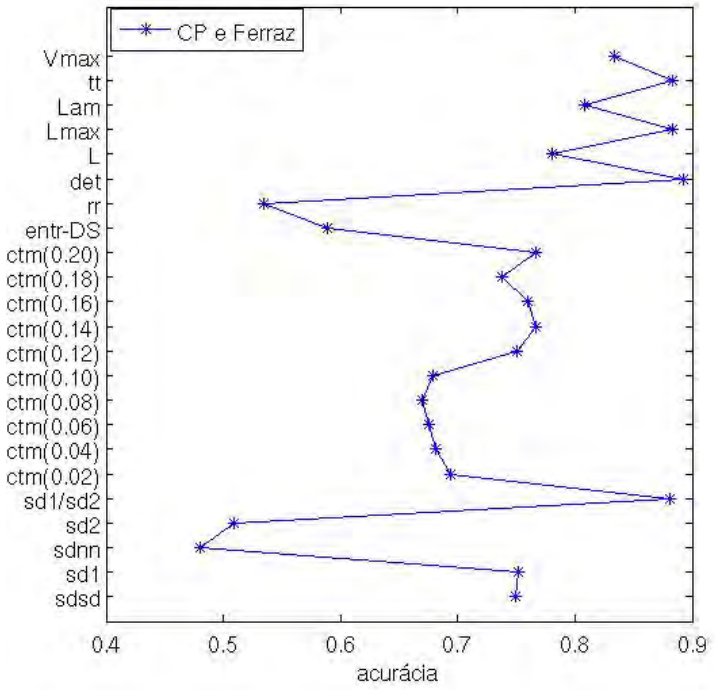

Figura 5.5 - Acurácias obtidas para cada uma das medidas apresentadas como entrada no SVM para o conjunto de séries temporais de vento zonal (CP e CF).

A Figura 5.6A apresenta o histograma de acurácias médias e a Figura 5.5B apresenta os valores de acurácia que cada medida obteve associada a outra medida fornecida como padrão de entrada. Os índices SD1/SD2, DET, L_{max} e LAM apresentam valores de acurácia $\geqslant 0,8$.

Observamos que os dois classificadores fornecem resultados semelhantes. Por exemplo, ambos apontam o valor do índice LAM é diferente entre as os dois conjuntos de séries temporais de vento. Aqui são mostrados apenas os resultados obtidos com o SVM.

5.5 Considerações sobre o Capítulo

Neste Capítulo foram apresentados a contextualização e os resultados obtidos com o segundo estudo de caso. Foram usadas séries temporais de vento zonal oriundas de banco de dados do INPE. Ao todo, foram obtidos 23 índices e foi realizada a comparação entre o conjunto de série de vento zonal de Cachoeira Paulista e o conjunto de séries de vento zonal da Estação Comandante Ferraz.

Diferentemente do primeiro estudo de caso, nesse contexto, os conjuntos de séries dis-

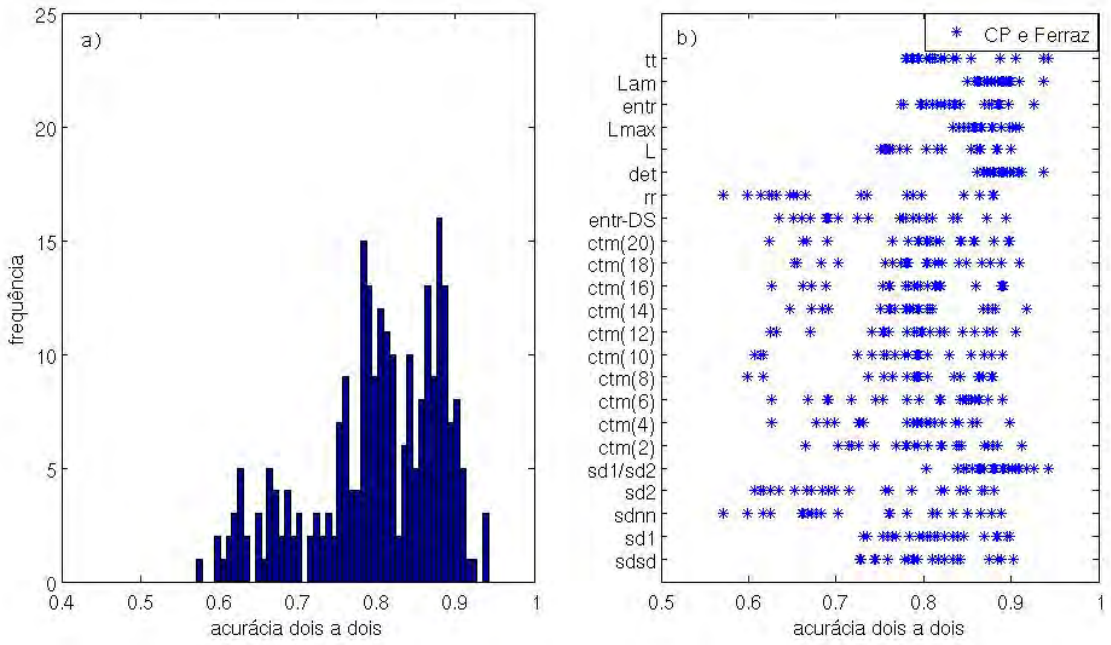

Figura 5.6 - Acurácias obtidas para as combinações dois a dois das 23 índices fornecidos como entrada para o SVM na comparação dos grupos Cachoeira Paulista e Comandante Ferraz.

tintos foram definidos pela localização do radar meteórico que os captou. Usualmente essas séries podem ser aproximadas por uma função senoidal, podendo apresentar diferentes periodicidades. Esperávamos portanto, que as medidas utilizadas (que não fazem parte dos estudos convencionais dessas séries) detectassem diferenças entre os dois conjuntos de dados.

A discriminação dos dois conjuntos foi detectada com o uso dos índices das medidas de quantificação de recorrência (RQA), obtendo um alto valor de acurácia média usando o J48 e o SVM. Os demais métodos não apresentaram-se adequados para discriminação desse conjunto de séries temporais.

6 CONCLUSÃO

O presente trabalho teve por objetivo caracterizar os conjuntos de sinais não lineares pelo uso associado de métodos de sistemas dinâmicos e técnicas de mineração de dados. Para tanto foi usado um conjunto de séries temporais de intervalos RR pré-classificadas pelo médico especialista em nove diferentes grupos.

Dentre as principais contribuições apresentadas nesse trabalho, destaca-se a análise realizada com o uso do filtro adaptativo nas séries de intervalos RR. O uso do filtro mostrou-se estatisticamente equivalente aos resultados obtidos com a utilização do método de filtragem convencional. Os parâmetros livres do filtro adaptativo foram escolhidos de forma empírica, de tal maneira que as séries filtradas pelo método adaptativo se aproximavam visualmente (conforme avaliação do médico especialista) das séries filtradas pelo método convencional. E, conforme alguns testes realizados, foi possível verificar o papel de cada parâmetro ajustado para o método adaptativo.

Como o desempenho do filtro adaptativo foi muito satisfatório, recomendamos este como uma ferramenta eficaz de pré-processamento das séries temporais de intervalos RR para análise da VFC (considerando apenas ritmo sinusal), auxiliando de maneira mais rápida a análise de um grande volume de dados. Salientamos que essa análise foi baseada no pré-processamento realizado pelo médico especialista e considerado nosso padrão-ouro.

Ainda em relação ao pré-processamento dos tacogramas, destacamos que o método adaptativo foi usado, além da maneira usual, como um detector dos intervalos RR, considerados não normais (conforme um critério pré-estabelecido e interessados na análise do ritmo sinusal) removendo-os da série total. Entretanto essa análise foi inicial, para alguns métodos de sistemas dinâmicos, pois exigiu um sistema de remoção de intervalos RR adequado. Ou seja, a análise na remoção desses intervalos RR deve ser específica considerando cada método não linear utilizado. Mais estudos serão necessários para uma análise final em relação à influência dos intervalos RR substituídos e/ou removidos nos conjuntos de tacogramas.

Os resultados obtidos com a discriminação das séries temporais de intervalos RR mostraram que não há um único índice não linear que detecta com desempenho satisfatório a diferença entre todos os grupos de tacogramas comparados. Ou seja, podemos dizer que, conforme os grupos comparados, houve valores de índices diferentes que detectaram diferenças entre os conjuntos de séries temporais.

Por exemplo, observou-se que entre os grupos CONT (crianças com peso normal) e COB (crianças com sobrepeso) nenhum índice atingiu uma acurácia média superior a $0,75$, indicando que esses índices não detectaram diferenças significativas na variabilidade desses conjuntos de séries de intervalos RR. Entretanto entre a comparação dos grupos RNN (recém-nascidos normais) e RNP (recém-nascidos prematuros) observou-se que houve índices que atingiram uma acurácia média superior a $0,75$ (SDNN, SD2 e SD1/SD2), evidenciando que existiram diferenças significativas entre os grupos, sendo esses índices capazes de detectá-las.

Por fim salientamos que, mesmo não realizando todas as comparações de grupos de tacogramas possíveis, detectamos indícios de existência de uma relação entre VFC e a faixa etária dos indivíduos. Isso pode estar relacionado à ontologia do sistema nervoso autônomo (com o ganho e perda das suas funções).

Os métodos de discriminação apresentados neste trabalho foram aplicados com sucesso no conjunto de séries temporais de intervalos RR. Verificamos que é possível aplicar estes métodos também em outros conjuntos de séries temporais. A seguir são listadas as principais conclusões obtidas da metodologia aplicada em cada estudo de caso:

Os índices que mais diferenciaram o conjunto de séries temporais de vento neutro foram relacionados ao RQA, que quantificam a recorrência presente nos sistemas. E, associando isso ao fato que essas séries temporais podem se diferenciar conforme a latitude em que foram captadas pelos radares meteóricos e que podem variar conforme a sazonalidade, ou seja, podemos dizer que há períodos "de recorrência" nas séries. Logo, esses resultados mostraram-se adequados.

6.1 Futuras perspectivas

Outras investigações podem ser sugeridas a partir deste trabalho em relação ao uso do filtro para pré-processamento dos tacogramas. Por exemplo, investigar métodos de otimização para ajuste automático dos parâmetros livres do filtro e verificar as alterações significativas na análise das séries temporais de intervalos RR. Outro ponto importante será verificar os efeitos da aplicação do filtro nas séries temporais.

Em relação ao conjunto de tacogramas, neste trabalho consideramos como referência a classificação dada pelo médico. Sabemos que essa classificação apresenta uma diversidade de casos influenciando também na diversidade da análise da VFC. Entretanto, poderíamos utilizar critérios para manter ou não determinados casos no

conjunto de séries, uniformizando assim os dados. Uma investigação futura relevante será analisar mais profundamente os grupos pré-classificados pelo especialista. Isso poderá ser feito tomando como base os métodos que quantificam a variabilidade de cada indivíduo, confrontando os valores obtidos com os métodos e a classificação do médico, com o objetivo de obtermos um grupo mais homogêneo em relação à VFC. Uma das vantagens em obtermos um grupo homogêneo em relação à VFC será a formação de um grupo de referência.

Por fim, como os métodos aplicados nos tacogramas apresentaram um resultado satisfatório, poderemos aplicar esses mesmos métodos em outros conjuntos de séries temporais que apresentarem um comportamento não linear.

REFERÊNCIAS BIBLIOGRÁFICAS

ACHARYA, U.; JOSEPH, K.; KANNATHAL, N.; LIM, C.; SURI, J. Heart rate variability: a review. **Med Bio Eng Comput**, v. 44, p. 1031–1051, 2006. 1, 2

AGUIRRE, L. A. **Introdução à identificação dos sistemas**: técnicas lineares e não lineares aplicadas a sistemas reais. Belo Horizonte: Editora UFMG, 2007. 3, 8, 9

ALVAREZ, D.; HORNERO, R.; GARCÍA, M.; CAMPO, F. del; ZAMARRON, C. Improving diagnostic ability of blood oxygen saturation from overnight pulse oximetry in obstructive sleep apnea detection by means of central tendency measure. **Artificial Intelligence in Medicine**, v. 41, p. 13–24, 2007. 4, 12

ANDREOLA, R. **Support vector machines na classificação de imagens hiperespectrais**. 130 p. Dissertação (Mestrado em Sensoriamento Remoto) — Universidade Federal do Rio Grande do Sul, Porto Alegre, 2009. 23

ANDRIOLI, V. F. **Variância e fluxo de momento devidos às ondas de gravidade na região MLT**. 183 p. Tese (Doutorado) — Instituto Nacional de Pesquisas Espaciais (INPE), São José dos Campos, 2012-08-15 2012. Disponível em: <http://urlib.net/sid.inpe.br/mtc-m19/2012/07.10.14.17>. 5, 94

ARAUJO, G. M. **Algoritmo para reconhecimento de características faciais baseado em filtros de correlação**. Dissertação (Mestrado) — Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2010. 28

Ayres-de-Campos, D.; COSTA-SANTOS, C.; BERNARDES, J. Sisporto multicentre validation study group. prediction of neonatal state by computer analysis of fetal heart rate tracings: the antepartum arm of the sisporto multicentre validation study. **Eur J Obstet Gynecol Reprod Biol**, v. 118, n. 1, p. 52–60, 2005. 44

BAGESTON, J. V. **Caracterização de ondas de gravidade mesosféricas na Estação Antártica Comandante Ferraz**. 176 p. Tese (Doutorado) — Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2009-12-16 2010. Disponível em: <http://urlib.net/sid.inpe.br/mtc-m19@80/2009/11.25.17.35>. 94

BAIM, D.; COLUCCI, W.; MONRAD, E.; SMITH, H.; WRIGHT, R.; LANOUE, A.; GAUTHIER, D.; RANSIL, B.; GROSSMAN, W.; BRAUNWALD, E. Survival

of patients with severe congestive heart failure treated with oral milrinone. **J Am Coli CardioI**, v. 7, p. 661–670, 1986. 44

BAKER, G. L.; GOLLUB, J. P. **Chaotic dynamics: an introduction**. Cambrigde: Cambrigde University Press, 1998. 2, 16

BERNARDES, J.; MOURA, C.; SA, J. P. de; LEITE, L. The Porto system for automated cardiotocographic signal analysis. **J Perinat Med**, v. 19(1-2), p. 61–65, 1991. 44

BIGGER, J.; FLEISS, L.; STEINMAN, R.; ROLNITZKY, L.; SCHNEIDER, W.; STEIN, P. Rr variability in healthy, middle-age persons compared with patients with chronic coronary heart disease or recent acute myocardial infarction. **Circulation**, v. 91, p. 1936–1943, 1995. 44

BITTENCOURT, M.; ROCHA, R.; Albanesi Filho, F. Cardiomiopatia hipertrófica. **Revista Brasileira de Cardiologia**, v. 23, n. 1, p. 17–24, 2010. 40

BRENNAN, M.; PALANISWAMI, M.; KAMEN, P. Do existing measures of poincaré plot geometry reflect nonlinear features of heart rate variability? **IEEE Transactions on Biomedical Engineering**, v. 48, n. 11, p. 1342–1347, 2001. 2, 9, 10, 55

CARNEY, R.; FREEDLAND, K. E.; STEIN, P.; MILLER, G.; STEINMEYER, B.; RICH, M. Heart rate variability and markers of inflammation and coagulation in depressed patients with coronary heart disease. **J Psychosom Res**, v. 62, n. 4, p. 463–467, 2007. 40

COHEN, M.; HUDSON, D.; DEEDWANIA, P. Applying continuous chaotic modeling to cardiac signal analysis. **IEEE Engineering in Medicine and Biology**, p. 97–102, 1996. 3, 13

CORTES, C.; VAPNIK, V. Support-vector networks. **Machine Learning**, v. 20, p. 273–297, 1995. 3, 28

COSTA, A.; CAMPOS, D. A. de; COSTA, F.; SANTOS, C.; BERNARDES, J. Prediction of neonatal acidemia by computer analysis of fetal heart rate and st event signals. **Am J Obstet Gynecol**, v. 201, p. 464.e1–6, 2009. 44

De Carvalho, J. L. A.; NOGUEIRA, O. S. A.; ROCHA, A. F.; NASCIMENTO, F. A. O.; NETO, J. S. Avaliação de métodos de interpolação do sinal de variabilidade da frequência cardíaca. In: XVIII CONGRESSO BRASILEIRO DE

ENGENHARIA BIOMéDICA, 1., 2002, São José dos Campos. **CBEB 2002 - A tecnologia da Vida - XVIII Congresso Brasileiro de Engenharia Biomédica**. São José dos Campos: UNIVAP, 2002. 3, 8

DOWNING, D.; CLARK, J. **Estatística Aplicada**. 3. ed. São Paulo: Saraiva, 2010. 117

DRISCOLL, W. Robustness of the ANOVA and Tukey-Kramer statistical tests. **Computers ind. Engng**, v. 31, p. 265–268, 1996. 64

ECKMANN, J. P.; KAMPHORST, S. O.; RUELLE, D. Recurrence plots of dynamical systems. **Europhysics Letters**, v. 7, p. 1035–1047, 1987. 3, 4, 17

FARIA, R. P. **Fundamentos de Astronomia**. 9. ed. São Paulo: Papirus, 2007. 93

FAWCETT, T. An introduction to ROC analysis. **Pattern Recognition Letters**, v. 27, p. 861–874, 2006. 120

FREITAS, U.; ROULIN, E.; MUIR, J.-F.; LETELLIER, C. Identifying chaos from heart rate: The right task? **Chaos**, v. 19, p. 1–4, 2009. 2, 35

FRITTS, D. C.; JANCHES, D.; IIMURA, H.; HOCKING, W. K.; BAGESTON, J. V.; LEME, N. M. P. Drake antarctic agile meteor radar first results: Configuration and comparison of mean and tidal wind and gravity wave momentum flux measurements with southern argentina agile meteor radar. **Journal of geophysical research**, v. 117, n. D02105, 2012. 5, 91, 94

GAMELIN, F. X.; BERTHOIN, S.; BOSQUET, L. Validity of the polar s810 heart rate monitor to measure rr intervals at rest. **Medicine & Science in Sports & Exercise**, v. 38, n. 5, p. 887–893, 2006. 42

GLASS, L. Introduction to controversial topics in nonlinear science: is the normal heart rate chaotic. **Chaos**, v. 19, p. 1–4, 2009. 2

GODOY, M. F. D.; TAKAKURA, I. T.; CORREA, P. R. Relevância da análise do comportamento dinâmico não linear (teoria do caos) como elemento prognóstico de morbidade e mortalidade em pacientes submetidos à cirurgia de revascularização miocárdica. **Arquivos de Ciência da Saúde**, v. 12, n. 4, p. 167–171, 2005. 42

GOLDBERGER, A. L.; AMARAL, L. A.; GLASS, L.; HAUSDORFF, J. M.; IVANOV, P. C.; MARK, R. G.; MIETUS, J. E.; MOODY, G. B.; PENG, C.-K.; STANLEY, H. E. Physiobank, physiotoolkit, and physionet: Components of a new

research resource for complex physiologic signals. **Circulation**, v. 101, p. e215–e220, 2000. 44, 62

GOLDSMITH, R. L.; BIGGER, J. T.; BLOOMFIELD, D. M.; KRUM, H.; STEINMAN, R. C.; SACKNER-BERNSTEIN, J.; PACKER, M. Long-term carvedilol therapy increases parasympathetic nervous system activity in chronic congestive heart failure. **The American Journal of Cardiology**, v. 80, p. 1101–1104, 1997. 62

GONCALVES, H.; BERNARDES, J.; ROCHA, A. P.; CAMPOS, D. A. de. Linear and nonlinear analysis of heart rate patterns associated with fetal behavioral states in the antepartum period. **Early Hum Dev**, v. 83, n. 9, p. 585–591, 2007. 44

GONCALVES, H.; ROCHA, A. P.; Ayres-de-Campos, D.; BERNARDES, J. Linear and nonlinear fetal heart rate analysis of normal and acidemic fetuses in the minutes preceding delivery. **Med Biol Eng Comput**, v. 44, n. 10, p. 847–855, 2006. 44

GONCALVES, H.; ROCHA, A. P.; CAMPOS, D. A. de; BERNARDES, J. Internal versus external intrapartum foetal heart rate monitoring: the effect on linear and nonlinear parameters. **Physiol Meas**, v. 27, p. 307–319, 2006. 44

GONCALVEZ, H.; HENRIQUES-COELHO, T.; ROCHA, A. P.; LOURENçO, A. P.; LEITE-MOREIRA, A.; BERNARDES, J. Comparison of different methods of heart rate entropy analysis during acute anoxia superimposed on a chronic rat model of pulmonary hypertension. **Medical Engineering & Physics**, v. 35, p. 559–568, 2013. 2

GOSHVARPOUR, A.; GOSHVARPOUR, A.; RAHATI, S. Analysis of lagged Poincaré plots in heart rate signals during meditation. **Digital Signal Processing**, v. 21, p. 208–214, 2011. 3, 9, 56

GUERRA, J. M. **Caracterização fina dos padrões de variabilidade do ECG para validação de modelos e aplicações em microgravidade**. 147 p. Dissertação (Mestrado) — Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2008-12-10 2008. Disponível em: <http://urlib.net/sid.inpe.br/mtc-m18@80/2009/01.29.13.50>. 41

GUO, R.; WANG, Y.; YAN, J.; YAN, H. Recurrence quantification analysis on pulse morphological changes in patients with coronary heart disease. **J Tradit Chin Med**, v. 32, n. 4, p. 571–577, 2012. 4

GUZZETTI, S.; BORRONI, E.; GARBELLI, P. E.; CERIANI, E.; BELLA, P. D.; MONTANO, N.; COGLIATI, C.; SOMERS, V. K.; MALLANI, A.; PORTA, A. Symbolic dynamics of heart rate variability: A probe to investigate cardiac autonomic modulation. **Circulation**, v. 112, p. 465–470, 2005. 3, 4

HAN, J.; KAMBER, M.; PEI, J. **Data Mining**: Concepts and techniques. 3. ed. San Francisco: Morgan Kaufmann Publishers, 2011. 24

HAYKIN, S. **Adaptive filter theory**. 3. ed. New Jersey: Prentice Hall Information and System Sciences Series, 1996. 45

ISLER, M. K. Y. Combining classical HRV indices with wavelet entropy measures improves to performance in diagnosing congestive heart failure. **Computers in Biology and Medicine**, v. 37, p. 1502–1510, 2007. 3, 9, 56

JAVORKA, M.; TURIANIKOVA, Z.; TONHAJZEROVA, I.; JAVORKA, K.; BAUMERT, M. The effect of orthostasis on recurrence quantification analysis of heart rate and blood pressure dynamics. **Physiological Measurement**, v. 30, n. 1, p. 29–41, 2009. 4

JEONG, J.; GORE, J.; PETERSON, B. A method for determinism in short time series, and its application to stationary eeg. **IEEE Transactions on Biomedical Engineering**, v. 49, n. 11, 2002. 3, 12

JOVIC, A.; BOGUNOVIC, N. Electrocardiogram analysis using a combination of statistical, geometric, and nonlinear heart rate variability features. **Artificial Intelligence in Medicine**, v. 51, n. 3, p. 175–186, 2011. 4

JUNG, Y.; PARK, S. Application of the autoregressive method to the spectral analysis of a flow signal. **Experiments in Fluids**, v. 31, p. 608–614, 2001. 3, 8

KANTZ, H.; SCHREIBER, T. **Nonlinear time series analysis**. Cambrigde: Cambrigde University Press, 2004. 2, 17

KAREMAKER, J. M.; BERECKI-GISOLF, J. 24-h blood pressure in space: The dark side of being an astronaut. **Respiratory Physiology & Neurobiology**, v. 169S, p. S55–S58, 2009. 2

KARLSSON, M.; HöRNSTEN, R.; RYDBERG, A.; WIKLUND, U. Automatic filtering of outliers in RR intervals before analysis of heart rate variability in holter recordings: a comparison with carefully edited data. **BioMedical Engineering OnLine**, v. 11, n. 2, 2012. 5

KARMAKAR, C. K.; KHANDOKER, A. H.; GUBBI, J.; PALANISWAMI, M. Complex correlation measure: a novel descriptor for poincare plot. **BioMedical Engineering Online**, v. 8, n. 17, 2009. 9

KEENAN, D. B.; GROSSMAN, P. Adaptive filtering of heart rate signals for an improved measure of cardiac autonomic control. **International Journal of Information and Communication Engineering**, v. 2, n. 1, p. 52–58, 2006. 5

KNIGHT, R. D. **Física: uma abordagem estratégica**. 2. ed. Porto Alegre: Bookman, 2009. 93, 94

KRUM, H.; JR, J. T. B.; GOLDSMITH, R. L.; PACKER, M. Effect of long-term digoxin therapy on autonomic function in patients with chronic heart failure. **Journal of the American College of Cardiology**, v. 25, p. 289–294, 1995. 62

KUUSELA, T. Heart rate variability (HRV) signal analysis. In: _____. Boca Raton: CRC Press, 2013. cap. Methodological aspects of heart rate variability analysis, p. 9–40. 36

LIJMER, J.; HUNINK, M.; DUNGEN, J. van den; LOONSTRA, J.; SMIT, A. Roc analysis of noninvasive tests for peripheral arterial disease. **Ultrasound in Med. & Biol**, v. 22, n. 4, p. 391–398, 1996. 121

LIMA, A. H. R. A.; FARAH, B. Q.; RODRIGUES, L. B. C. C.; MIRANDA, A. S.; RODRIGUES, S. L. C.; CORREIA, M. de A.; FILHO, D. C. S.; FORJAZ, C. L. M.; PRADO, W. L.; WOLOSKER, N.; RITTI-DIAS, R. M. Low-intensity resistance exercise does not affect cardiac autonomic modulation in patients with peripheral artery disease. **Clinics**, v. 68, n. 5, p. 632–637, 2013. 4

LING, L. Y. **Uso combinado de métodos de dinâmica não linear e redes neurais na avaliação da variabilidade da frequência cardíaca em diferentes situações clínicas**. 112 p. Dissertação (Mestrado em Computação Aplicada) — Instituto Nacional de Pesquisas Espaciais (INPE), São José dos Campos, 2009. 21

LOGIER, R.; DE JONCKHEEREonckheere, J.; DASSONNEVILLE, A. An efficient algorithm for R-R intervals series filtering. In: IEEE, 26., 2004, San Francisco, CA. **Engineering in Medicine and Biology Society**. San Francisco, CA: IEEE, 2004. 5

LORENA, A. C.; CARVALHO, A. C. P. L. F. de. Uma introdução às support vector machines. **Revista de Informática Teórica e Aplicada**, v. 14, n. 2, p. 43–67, 2007. 23, 28

MANTEL, R.; GEIJN, H. P. van; CARON, F. J.; SWARTJES, J. M.; WOERDEN, E. E. van; JONGSMA, H. W. Computer analysis of antepartum fetal heart rate: 1. baseline determination. **Int J Biomed Comput**, v. 25, p. 261–272, 1990. 44

MARTINELLI, L.; SANTOS, T. R. M.; FP, F. F.; BAUER, M.; MACHADO, A.; SUNDARESAN, A. Effect of microgravity on immune cell viability and proliferation: simulation using 3-d clinostat. **IEEE Engineering in Medicine and Biology Magazine**, v. 28, n. 4, p. 85–90, 2009. 40

MARTINEZ, E.; LOUZADA-NETO, F.; PEREIRA, B. A curva ROC para testes diagnosticos. **Cadernos Saúde Coletiva**, v. 11, n. 1, p. 7–31, 2003. 120

MARWAN, N.; ROMANO, M. C.; THIEL, M.; KURTHS, J. Recurrence plots for the analysis of complex systems. **Physics Reports**, v. 438, p. 237–329, 2007. xv, 3, 4, 17, 19, 20, 21, 22

MARWAN, N.; WESSE, N.; MEYERFELDT, U.; SCHIRDEWAN, A.; KURTHS, J. Recurrence-plot-based measures of complexity and their application to heart-rate-variability data. **Physical Review E**, v. 66, p. 026702/1–026702/8, 2002. 3, 4, 20

MAZON, J.; GASTALDI, A.; DI SACCI, T.; COZZA, I.; DUTRA, S.; SOUZA, H. Effects of training periodization on cardiac autonomic modulation and endogenous stress markers in voleyball players. **Scand J Med Sports**, v. 23, p. 114–120, 2013. 1

MELILLO, P.; BRACALE, M.; PECCHIA, L. Nonlinear heart rate variability features for real-life stress detection. case study: students under stress due to university examination. **BioMedical Engineering OnLine**, v. 10, n. 96, 2011. 3

MEYER, P. L. **Probabilidade: Aplicações à Estatística**. 1. ed. Rio de Janeiro: Livros Técnicos e Científicos Editora S.A, 1969. 117, 118

MONTEIRO, L. H. A. **Sistemas dinâmicos**. São Paulo: Editora Livraria da Física, 2006. 8

MOUROT, L.; BOUHADDI, M.; PERREY, S.; ROUILLON, J. D.; REGNARD, J. Quantitative Poincaré plot analysis of heart rate variability: effect of endurance

training. **European Journal of Applied Physiology**, v. 91, p. 79–87, 2004. 3, 9, 55

NAECK, R. **Evaluation de l'adaptation à la ventilation non invasive chez des patients atteints d'insuffisance respiratoire chronique**. Tese (Doutorado) — Université de Rouen, 2011. 43

NATIONAL TAIWAN UNIVERSITY. **LIBSVM: A Library for Support Vector Machines**. Taiwan, abril 2012. Disponível em: <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>. 28, 30, 68

NUNAN, D.; DONOVAN, G.; JAKOVLJEVIC, D. G.; HODGES, L. D.; SANDERCOCK, G. R.; BRODIE, D. A. Validity and reliability of short-term heart-rate variability from the polar s810. **Med Sci Sports Exerc**, v. 41, n. 1, p. 243–250, 2009. 42

OLIVEIRA, R.; COSTA, M. V. da; PEDRO, R.; POLITO, M.; AVELAR, A.; CYRINO, E.; NAKAMURA, F. Acute cardiac autonomic responses after a bout of resistance exercise. **Science & Sports**, v. 27, n. 6, p. 357–364, 2012. 1

PAGANI, M. Heart rate variability and autonomic diabetic neuropathy. **Diabetes Nutrition & Metabolism**, v. 13, p. 341–346, 2000. 1

PETKOVIC, D.; COJBASIC, Z. Adaptive neuro-fuzzy estimation of autonomic nervous system parameters effect on heart rate variability. **Neural Comput & Applic**, v. 21, p. 2065–2070, 2012. 9

PIQUEIRA, J. R. C.; MATTOS, S. H. V. de. Note on LMC complexity measure. **Ecological Modelling**, v. 222, p. 3603–3604, 2011. 3, 4, 15

PISKORSKI, J.; GUZIK, P. Geometry of the Poincaré plot of RR intervals and its asymmetry in healthy adults. **Physiological Measurement**, v. 28, p. 287–300, 2007. 3, 10

POINCARÉ, H. **Les méthodes nouvelles de la mécanique celeste**. Paris: Gauthier-Villars, 1892. 16

POOL, R. Is it healthy to be chaotic? **Science**, v. 243, p. 604–607, 1989. 2, 35

QUINLAN, J. Induction of decision trees. **Machine Learning**, v. 1, p. 81–106, 1986. 24

RAMÍREZ-ROJAS, A.; FLORES-MÁRQUEZ, E. Order parameter analysis of seismicity of the Mexican Pacific coast. **Physica A: Statistical Mechanics and its Applications**, v. 392, n. 10, p. 2507–2512, 2013. 1

REDDY, L. R. G.; KUNTAMALLA, S. Analysis of degree of nonlinearity and stochastic nature of HRV signal during meditation using delay vector variance method. In: **33rd Annual International Conference of the IEEE Engineering-in-Medicine-and-Biology-Society (EMBS)**. Boston: IEEE, 2011. p. 2720–2723. 35

ROCHA, R. S. da. **Árvores de Decisão na Classificação de Objetos Astronômicos a partir de Parâmetros Fotométricos (Monografia de Qualificação)**. São José dos Campos: INPE, 2008. 24

RODRIGUEZ-ALVARES, M.; TAHOCES, P.; CADARSO-SUAREZ, C.; LADO, M. Comparative study of roc regression techniques - applications for the computer-aided diagnostic system in breast cancer detection. **Computational Statistics and Data Analysis**, v. 55, p. 888–902, 2011. 121

ROY, B.; CHOUDHURI, R.; PANDEY, A.; BANDOPADHYAY, S.; SARANGI, S.; GHATAK, S. K. Effect of rotating acoustic stimulus on heart rate variability in healthy adults. **The Open Neurology Journal**, v. 6, p. 71–77, 2012. 3

RUGGIERO, M. A. G.; LOPES, V. L. da R. **Cálculo Numérico: Aspectos Teóricos e Computacionais**. 2. ed. São Paulo: Pearson, 2004. 95

SAFAVIAN, S. R.; LANDGREBE, D. A survey of decision tree classifier methodology. **IEEE Transactions on Systems, Man, and Cybernetics**, v. 21, n. 3, p. 660–674, 1991. 3, 23, 24, 25

SELIG, F. A.; TONOLLI, E. R.; SILVA, E. V. C. M. da; GODOY, M. F. de. Heart rate variability in preterm and term neonates. **Arquivos Brasileiros de Cardiologia**, v. 96, n. 6, p. 443–449, 2011. 1, 42, 62

SMALE, S. Differential dynamical systems. **Bull. Amer. Math. Soc.**, v. 73, p. 747–817, 1967. 15

SOUZA, E. G. **Caracterização de sistemas dinâmicos através de gráficos de recorrência**. 104 p. Dissertação (Mestrado em Física) — Universidade Federal do Paraná, Curitiba, 2008. 16, 17, 23

STATSDIRECT STATISTICAL SOFTWARE. 2013. Disponível em: <http://www.statsdirect.com/>. Acesso em: 11 junho 2013. 119

STEIN, P.; EHSANI, A.; DOMITROVICH, P.; KLEIGER, R.; ROTTMAN, J. The effect of exercise training on heart rate variability in healthy older adults. **American Heart Journal**, v. 138, p. 567–576, 1999. 44

SUNKARIA, R. K. Recent trends in nonlinear methods of HRV analysis: A review. **World Academy of Science, Engineering and Technology**, v. 75, p. 566–571, 2011. 1

SZALBERG, S. L. Book review: C4.5: Programs for machine learnings by J. Ross Quinlan, Morgan Kaufmann Publishers Inc., 1993. **Machine Learning**, v. 16, p. 235–240, 1994. 24

Task Force of the European Society of Cardiology the North American Society of Pacing Electrophysiology. **Circulation**, v. 93, p. 1043–1065, 1996. 1, 2, 43

TRULLA, L. L.; GIULIANI, A.; ZBILUT, J. P.; JR., C. L. W. Recurrence quantification analysis of the logistic equation with transients. **Physics Letters A**, v. 223, p. 255–260, 1996. 1, 20

VANDERLEI, L.; SILVA, R.; PASTRE, C.; AZEVEDO, F.; GODOY, M. Comparison of the Polar S810i monitor and the ecg for the analysis of heart rate variability in the time and frequency domains. **Brazilian Journal of Medical and Biological Research**, v. 41, n. 10, p. 854–859, 2008. 42

VANDERLEI, L. C. M.; PASTRE, C. M.; HOSHI, R. A.; CARVALHO, T. D. de; GODOY, M. F. de. Noções básicas de variabilidade da frequência cardíaca e sua aplicabilidade clínica. **Rev Bras Cir Cardiovasc**, v. 24, n. 2, 2009. 1, 9, 36, 40

VANDERLEI, L. C. M.; PASTRE, C. M.; JúNIOR, I. F. F.; GODOY, M. F. de. Índices geométricos de variabilidade da frequência cardíaca em crianças obesas e eutróficas. **Arquivos Brasileiros de Cardiologia**, v. 95, n. 1, p. 35–40, 2010. 42

VAPNIK, V. **Statistical Learning Theory**. New York: Wiley, 1998. 29

WANG, J.-Y. **Application of Support Vector Machines in Bioinformatics**. 65 p. Master of Science (Computer Science and Information Engineering) — National Taiwan University, Taiwan, 2002. 28, 29, 30

WESSEL, N.; MALBERG, H.; ZIEHMANN, C.; VOSS, H. U.; SCHIRDEWAN, A.; MEYERFELDT, U.; KURTHS, J. Nonlinear analysis of complex phenomena in cardiological data. **Herzschr Elektrophys**, v. 11, p. 159–173, 2000. 5, 45

WITTEN, I. H.; FRANK, E. **Data Mining: practical machine learning tools and techniques**. 2. ed. Boston: Elsevier Inc, 2005. xxi, 2, 26, 27, 68

WONGKHAM, S.; BOONLA, C.; KONGKHAM, S.; WONGKHAM, C.; BHUDHISAWASDI, V.; SRIPA, B. Serum total sialic acid in cholangiocarcinoma patients: an roc curve analysis. **Clinical Biochemistry**, v. 34, p. 537–541, 2001. 121

WOO, M. A.; STEVENSON, W. G.; MOSER, D. K.; MIDDLEKAUFF, H. R. Complex heart rate variability and serum norepinephrine levels in patients with advanced heart failure. **Journal of the American College Cardiology**, v. 23, p. 565–569, 1994. 3

WRASSE, C. M. **Estudos de geração e propagação de ondas de gravidade atmosféricas**. Tese (Doutorado) — Instituto Nacional de Pesquisas Espaciais, 2004. xix, 91, 92, 93, 94

XU, H.; LOHR, J.; GREINER, M. The selection of elisa cut-off points for testing antibody to newcastle disease by two-graph receiver operating characteristic (tg-roc) analysis. **Journal of Immunological Methods**, v. 208, p. 61–64, 1997. 121

YU, S.; LEE, M. Conditional mutual information-based feature selection for congestive heart failure recognition using heart rate variability. **Computer Methods and Programs in Biomedicine**, v. 108, n. 1, p. 299–309, 2012. 4

APÊNDICE B - ANÁLISE ESTATÍSTICA

Nesse Apêndice são descritos os métodos estatísticos usados na análise dos dados apresentados na Seção 4.1. Dentre os objetivos da utilização da análise estatística listados na Seção 4.1 podemos destacar: avaliar a equivalência estatística dos métodos de filtragem empregados e analisar a influência do tamanho das séries temporais de intervalos RR. Entretanto, aqui são apenas relatados os métodos e não suas aplicabilidades.

A.1 Teste t de Student

O Teste t de *Student* é um teste de hipótese que faz uso de conceitos estatísticos para aceitar ou rejeitar uma hipótese nula[1] quando essa estatística de teste segue uma distribuição t de *Student*[2].

Após formular a hipótese nula, o valor de t é obtido e aplicado à função densidade de probabilidade da distribuição t de *Student*, onde é medida a área abaixo da função para valores maiores ou iguais a t. Essa área será a probabilidade da média dessa amostra ter apresentado o valor observado ou não. Se a probabilidade de ocorrência deste resultado ter ocorrido for muito pequena, o resultado é estatisticamente relevante. Essa probabilidade também é chamada de $p-$valor ou valor p.

Estabelecendo-se um valor de corte para o $p-$valor, podemos rejeitar ou aceitar H_0. Se $p-$valor for menor que o ponto de corte, H_0 é rejeitada. Caso contrário não é rejeitada. É usual o ponto de corte ser considerado 5% ou $0,05$, ou seja, se a área abaixo da função densidade de probabilidade da distribuição t for menor, diz-se que a hipótese nula é rejeitada com nível de confiança de 95%[3].

Para calcular o valor t, considera-se que as variâncias das duas amostras analisadas (séries filtradas com método adaptativo e com método convencional) são iguais e:

$$t = \frac{\bar{x_1} - \bar{x_2}}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \tag{A.1}$$

[1] Hipótese nula é a hipótese que será testada no teste estatístico, denominada como H_0. A outra única possibilidade é que H_0 seja falsa, sendo que a hipótese que afirma que H_0 é falsa, é denominada como *hipótese alternativa* ou H_1 (DOWNING; CLARK, 2010).

[2] Uma distribuição t é uma distribuição normal, entretanto a variância não é conhecida (MEYER, 1969).

[3] É comum fixar-se em 95% a probabilidade da média da amostra estar no intervalo, denominado intervalo de confiança e 95% é o nível de confiança (DOWNING; CLARK, 2010).

sendo

$$s^2 = \frac{\sum_{j=1}^{n_1}(x_j - \bar{x_1})^2 + \sum_{i=1}^{n_2}(x_i - \bar{x_2})^2}{n_1 + n_2 - 2} \tag{A.2}$$

onde $\bar{x_1}$ e $\bar{x_2}$ são as médias das amostras, s^2 é a variância amostral coletiva, n_1 e n_2 são os tamanhos das amostras e t é um quantil[4] com $n_1 + n_2 - 2$ graus de liberdade.

Quando as amostras não seguem uma distribuição normal[5], o teste alternativo é o teste não paramétrico **Teste U** ou **Teste Mann-Whitney**. Este teste é definido como:

$$U = n_1 n_2 + \frac{n_2(n_2+1)}{2} - \sum_{i-n_1+1}^{n_2} R_i \tag{A.3}$$

onde as amostras de tamanho n_1 e n_2 são agrupadas e R_i são as classificações (*ranks*).

A.2 Coeficiente de correlação de Pearson

O coeficiente de correlação de Pearson (r) mede o grau de correlação entre duas variáveis (x e y). A estimativa empregada para r (MEYER, 1969) é:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{A.4}$$

sendo $\bar{x}$ e $\bar{y}$ as médias das variáveis x e y e n o tamanho das amostras.

O coeficiente r assume valores entre -1 e 1:

- $r = 1$: a correlação é perfeita e positiva entre as duas variáveis;
- $r = -1$: a correlação é perfeita e negativa entre as duas variáveis, ou seja, se uma aumenta a outra diminui;
- $r = 0$: não há correlação entre as duas variáveis, mas pode existir uma dependência não linear não detectada por esse coeficiente.

Caso as variáveis não apresentem uma distribuição normal, calcula-se o **coeficiente de Spearman**. Para o cálculo do coeficiente de Spearman, há um classificação dos pares das variáveis (x e y) cada contendo n observações (STATSDIRECT STATISTICAL

[4]Quantis são porções de mesma quantidade ou tamanho de um todo. Por exemplo, dividir uma série temporal em quatro séries menores de mesmo tamanho, haverá quatro quantis. Cada será denominado de quartil.

[5]Para saber se a variável apresenta um distribuição normal é calculado o teste da normalidade.

SOFTWARE, 2013):

$$r = \frac{\sum_{i=1}^{n} R(x_i)R(y_i) - n\left(\frac{n+1}{2}\right)^2}{\left(\sum_{i=1}^{n} R(x_i)^2 - n\left(\frac{n+1}{2}\right)^2\right)^{0.5}\left(\sum_{i=1}^{n} R(y_i)^2 - n\left(\frac{n+1}{2}\right)^2\right)^{0.5}} \tag{A.5}$$

sendo $R(x)$ e $R(y)$ as classificações dos pares (x e y).

A.3 Análise de Variância - ANOVA

A análise de Variância, denominada ANOVA é um método de teste de hipótese aplicado com a finalidade de comparar as médias de vários grupos distintos. O teste ANOVA (*one way*) pode ser considerado uma generalização do teste t de duas amostras. A estatística F compara a variabilidade entre os grupos com a variabilidade dentro dos grupos:

$$F = \frac{MST}{MSE} \tag{A.6}$$

sendo

$$MST = \frac{\sum_{i=1}^{k}(T_i^2/n_i) - G^2/n}{k-1} \tag{A.7}$$

$$MSE = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}^2 - \sum_{i=1}^{k}(T_i^2/n_i)}{n-k} \tag{A.8}$$

onde F é a razão de variância de teste total, MST é a média entre os grupos (tratamento/grupos) e MSE é a média devido ao erro (dentro dos grupos), Y_{ij} é uma observação, T_i é o grupo total, G é total das observações, n_i é o número do grupo i e n é o número total de observações (STATSDIRECT STATISTICAL SOFTWARE, 2013).

Algumas observações a respeito desse teste:

- cada observação é independente das demais,
- cada tratamento possui distribuição normal,
- todas as distribuições tem a mesma variância.

Quando o $p-$valor obtido pelo teste ANOVA for $p \geqslant 0,05$, a hipótese nula é aceita e quando $p < 0,05$ implica rejeitar H_0 já que pelo menos uma das médias é diferente do grupo. Para descobrir qual média se diferencia das demais é aplicado a **Comparação Múltipla de Tukey**.

O teste de Tukey é utilizado quando se deseja comparar todos os pares de médias de r populações, adotando-se um único nível de significância. O teste consiste em calcular um valor acima do qual a diferença entre duas médias amostrais (em valor absoluto) é significativamente diferente de zero.

A.4 Análise da curva ROC

A análise pela curva ROC ("Receiver Operating Characteristic Curve") é uma técnica de classificação amplamente baseada na teoria de detecção do sinal para representar a troca entre as taxas de sucesso e as taxas de falsa classificação dada por uma determinada variável, podendo ser usada para avaliar a aplicabilidade de métodos diagnósticos (FAWCETT, 2006).

Inicialmente, segundo Fawcett (2006), considera-se o problema de classificação usando duas classes. Cada caso i é mapeado para um elemento do conjunto de classe positiva (P) (possuir uma determinada característica, por exemplo, estar doente) ou negativa (N) (não possuir uma determinada característica, por exemplo, não estar doente). Para distinguir entre a classe atual e a classe preditiva são usados os rótulos SIM (tem a característica, conforme um determinado ponto de corte estabelecido) e NÃO (não possui a característica conforme um determinado ponto de corte estabelecido) para as classes preditas produzidas pelo modelo (ver Tabela A.1).

Tabela A.1 - Exemplo de tabela de contingência ou matriz de confusão para classificar modelos pela análise ROC.

	P	N	Total
Sim	TP	FP	TP+FP
Não	FN	TN	FN+VN
Total	TP+FN	FP+TN	TP+FP+FN+VN

considerando TP = verdadeiro positivo - número de pacientes doentes classificados pelo teste; FP = falsos positivos - número de pacientes não doentes não classificados pelo teste; FN = falsos negativos - número de pacientes doentes não classificados pelo teste; TN = verdadeiros negativos - número de pacientes não doentes classificados corretamente pelo teste.

Nas pesquisas médicas e biológicas, a análise pela curva ROC é uma prática muito comum, principalmente quando se deseja verificar o desempenho de um teste diagnóstico (MARTINEZ et al., 2003). Essa técnica já foi aplicada em inúmeros estudos,

podendo-se citar como exemplo, para avaliar a acurácia do diagnóstico de testes não invasivos selecionados na doença arterial periférica (LIJMER et al., 1996), avaliar sistemas de diagnósticos em detecção de câncer mamário (RODRIGUEZ-ALVARES et al., 2011), determinar a utilidade diagnóstica do nível sérico de ácido siálico total em pacientes com colangiocarcinoma (câncer das vias biliares) (WONGKHAM et al., 2001) e seleção de pontos de cortes para testes de anticorpos para doença de Newcastle (doença viral altamente contagiosa que afeta aves domésticas e selvagens) (XU et al., 1997) entre outros.

A partir da Tabela A.1 algumas informações podem ser obtidas:

(i) a sensibilidade (S = VP/(VP + FN)) que pode ser definida como a probabilidade do teste fornecer um resultado positivo, dado que o indivíduo realmente é positivo - portador da característica desejada;

(ii) a especificidade (E = VN/(FP + VN)) que pode ser definida como a probabilidade do teste fornecer um resultado negativo, dado que o indivíduo realmente é negativo - não portador da característica desejada;

(iii) o valor preditivo positivo (VPP = VP/(VP + FP)) que pode ser definido como a porcentagem de pacientes com resultado positivo que são doentes;

(iv) o valor preditivo negativo (VPN = VN/(FN + VN)) definido como resultados negativos que não estão doentes;

(v) e a razão de verossimilhança (LRP = sensibilidade/(1-especificidade)) que é correspondente à relação entre a probabilidade de um teste diagnóstico dar positivo em quem tem a doença (sensibilidade) e a probabilidade desse mesmo teste dar positivo em quem não tem a doença (1-especificadade).

PUBLICAÇÕES TÉCNICO-CIENTÍFICAS EDITADAS PELO INPE

Teses e Dissertações (TDI)

Teses e Dissertações apresentadas nos Cursos de Pós-Graduação do INPE.

Manuais Técnicos (MAN)

São publicações de caráter técnico que incluem normas, procedimentos, instruções e orientações.

Notas Técnico-Científicas (NTC)

Incluem resultados preliminares de pesquisa, descrição de equipamentos, descrição e ou documentação de programas de computador, descrição de sistemas e experimentos, apresentação de testes, dados, atlas, e documentação de projetos de engenharia.

Relatórios de Pesquisa (RPQ)

Reportam resultados ou progressos de pesquisas tanto de natureza técnica quanto científica, cujo nível seja compatível com o de uma publicação em periódico nacional ou internacional.

Propostas e Relatórios de Projetos (PRP)

São propostas de projetos técnico-científicos e relatórios de acompanhamento de projetos, atividades e convênios.

Publicações Didáticas (PUD)

Incluem apostilas, notas de aula e manuais didáticos.

Publicações Seriadas

São os seriados técnico-científicos: boletins, periódicos, anuários e anais de eventos (simpósios e congressos). Constam destas publicações o Internacional Standard Serial Number (ISSN), que é um código único e definitivo para identificação de títulos de seriados.

Programas de Computador (PDC)

São a seqüência de instruções ou códigos, expressos em uma linguagem de programação compilada ou interpretada, a ser executada por um computador para alcançar um determinado objetivo. Aceitam-se tanto programas fonte quanto os executáveis.

Pré-publicações (PRE)

Todos os artigos publicados em periódicos, anais e como capítulos de livros.