UNIVERSIDADE DE SÃO PAULO
ESCOLA DE ARTES, CIÊNCIAS E HUMANIDADES
PROGRAMA DE PÓS-GRADUAÇÃO EM SISTEMAS DE INFORMAÇÃO

HENRIQUE DOS SANTOS PASSOS

Ensemble de técnicas de representação simbólica para reconhecimento biométrico baseado em sinais de ECG

São Paulo
2018

HENRIQUE DOS SANTOS PASSOS

Ensemble de técnicas de representação simbólica para reconhecimento biométrico baseado em sinais de ECG

Dissertação apresentada à Escola de Artes, Ciências e Humanidades da Universidade de São Paulo para obtenção do título de Mestre em Ciências pelo Programa de Pós-graduação em Sistemas de Informação.

Área de concentração: Metodologia e Técnicas da Computação

Versão corrigida contendo as alterações solicitadas pela comissão julgadora em 19 de abril de 2018. A versão original encontra-se em acervo reservado na Biblioteca da EACH-USP e na Biblioteca Digital de Teses e Dissertações da USP (BDTD), de acordo com a Resolução CoPGr 6018, de 13 de outubro de 2011.

Orientador: Prof. Dr. Clodoaldo Aparecido de Moraes Lima

São Paulo

2018

CATALOGAÇÃO-NA-PUBLICAÇÃO
(Universidade de São Paulo. Escola de Artes, Ciências e Humanidades. Biblioteca)
CRB 8 - 4936

Passos, Henrique dos Santos
Ensemble de técnicas de representação simbólica para reconhecimento biométrico baseado em sinais de ECG / Henrique dos Santos Passos ; orientador, Clodoaldo Aparecido de Moraes Lima – 2018.
103 f. : il

Dissertação (Mestrado em Ciências) - Programa de Pós-Graduação em Sistemas de Informação, Escola de Artes, Ciências e Humanidades, Universidade de São Paulo.
Versão corrigida

1. Inteligência artificial. 2. Biometria. 3. Eletrocardiografia. I. Lima, Clodoaldo Aparecido de Moraes, orient. II. Título

CDD 22.ed.– 006.3

Dissertação de autoria de Henrique dos Santos Passos, sob o título **"Ensemble de técnicas de representação simbólica para reconhecimento biométrico baseado em sinais de ECG"**, apresentada à Escola de Artes, Ciências e Humanidades da Universidade de São Paulo, para obtenção do título de Mestre em Ciências pelo Programa de Pós-graduação em Sistemas de Informação, na área de concentração Metodologia e Técnicas da Computação, aprovada em 19 de abril de 2018 pela comissão julgadora constituída pelos doutores:

Prof. Dr. Clodoaldo Aparecido de Moraes Lima

Instituição: Escola de Artes, Ciências e Humanidades (EACH), Universidade de São Paulo (USP)

Presidente

Prof. Dr. Romis Ribeiro de Faissol Attux

Instituição: Universidade Estadual de Campinas (Unicamp)

Prof. Dr. Maurício Marengoni

Instituição: Universidade Presbiteriana Mackenzie

Prof. Dr. Moises Vidal Ribeiro

Instituição: Universidade Federal de Juiz de Fora (UFJF)

Agradecimentos

Agradeço a Deus. Agradeço aos meus familiares por todo apoio as minhas pesquisas. Agradeço ao meu orientador o professor Clodoaldo Aparecido de Moraes Lima pelas orientações, conselhos, compreensões e ajudaras fornecidas durante o desenvolvimento dessa dissertação, é uma pessoa a se espelhar. Agradeço a professora Sarajane Marques Peres pela sua atenção e grande auxilio em publicações.

"Pois, que aproveitaria ao homem ganhar todo o mundo e perder a sua alma?"

(Marcos 8:36)

Resumo

PASSOS, Henrique dos Santos. **Ensemble de técnicas de representação simbólica para reconhecimento biométrico baseado em sinais de ECG**. 2018. 103 f. Dissertação (Mestrado em Ciências) – Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo, 2018.

Métodos de identificação de pessoas sempre foram muito importantes para toda a sociedade. Atualmente, as pesquisas em biometria vêm sendo amplamente incentivadas por diversos setores da indústria mundial com o objetivo de melhorar ou substituir os atuais sistemas de segurança e de identificação de pessoas. O campo da biometria abarca uma grande variedade de tecnologias usadas para identificar e verificar a identidade de uma pessoa por meio da mensuração e análise de diversas características físicas e/ou comportamentais do ser humano. Diversas modalidades biométricas têm sido propostas para reconhecimento de pessoas, como impressão digital, íris, face e fala. Estas modalidades biométricas possuem características distintas em termos de desempenho, mensurabilidade e aceitabilidade. Uma questão a ser considerada com a aplicação biométrica é sua robustez a ataques por circunvenção, repetição e ofuscação. Esses ataques estão se tornando cada vez mais frequentes e questionamentos estão sendo levantados a respeito dos níveis de segurança das formas de reconhecimento. Sinais biomédicos como eletrocardiograma (ECG), eletroencefalograma (EEG) e eletromiograma (EMG) têm sido cada vez mais estudados e aplicados ao reconhecimento biométrico. Em específico, os sinais de ECG têm sido largamente adotados para o reconhecimento biométrico em diversos trabalhos. Por outro lado, análise de séries temporais tem sido usada com sucesso em muitas diferentes aplicações para identificar padrões temporais nos dados. Embora dinâmica simples possa ser observada com ferramentas analíticas tradicionais tais como transformada de fourier, transformada wavelet, a representação simbólica pode melhorar a análise de processos que são complexos e possivelmente caótico. Além disso, representação simbólica pode também reduzir a sensibilidade a ruído e melhorar bastante a eficiência computacional. No entanto, existem aspectos estruturais e paramétricos de projeto que podem conduzir a uma degradação de desempenho. Na ausência de uma metodologia sistemática e de baixo custo para a proposição de técnicas de representação simbólicas otimamente especificadas, os comitês de máquinas, mais especificamente ensemble, se apresentam como alternativas promissoras. Neste estudo, os componentes do ensemble, que correspondem as técnicas de representação simbólicas, e seus respectivos parâmetros foram selecionados via algoritmos evolutivos. O objetivo é explorar conjuntamente potencialidades advindas das técnicas de representação simbólicas e comitê de máquinas para reconhecimento biométrico baseado em sinais de ECG. Resultados experimentais conduzidos sobre dois conjuntos de dados disponíveis publicamente indicam que a abordagem proposta pode melhorar o desempenho do reconhecimento quando comparada com as técnicas tradicionais.

Palavras-chaves: Sistemas Biométricos; Sinais Biomédicos; Eletrocardiograma; Representação Simbólica; Comitê de Máquinas; Algoritmos Evolutivos.

Abstract

PASSOS, Henrique dos Santos. **Ensemble of symbolic representation techniques for biometric recognition based on ECG signals**. 2018. 103 p. Dissertation (Master of Science) – School of Arts, Sciences and Humanities, University of São Paulo, São Paulo, 2018.

Identification people methods have been very important for the whole society. Currently, research on biometrics have been widely encouraged by various sectors of the industry worldwide in order to improve or replace existing security systems and people identification. The field of biometrics includes a variety of technologies used to identify or verify the identity of a person by measuring and analyzing various physical and/or behavioral aspects of the human being. Several biometric methods have been proposed for recognition of people, such as fingerprint, iris, face and speech. These biometric modalities have different characteristics in terms of performance, measurability and acceptability. One issue to be considered with the biometric application in the real world is its robustness to attacks by circumvention, repetition and obfuscation. These attacks are becoming more frequent and more questions are being raised about the levels of security that this technology can offer. Biomedical signals such as electrocardiogram (ECG), electroencephalogram (EEG) and electromyogram (EMG) have been increasingly studied and applied to biometric recognition. Specifically, ECG signals have been widely adopted for biometric recognition in various works. On the other hand, time series analysis has been used successfully in many different applications to identify temporal patterns in the data. Although simple dynamics can be observed with traditional analytical tools such as fourier transform, wavelet transform, the symbolic representation can improve the analysis of processes that are complex and possibly chaotic. In addition, symbolic representation can also reduce noise sensitivity and greatly improve computational efficiency. However, there are structural and parametric design aspects that can lead to performance degradation. In the absence of a systematic and inexpensive methodology for proposing optimally specified symbolic representation techniques, machine committees, more specifically ensemble, present themselves as promising alternatives. In this study, the components of the committee, which correspond to techniques of symbolic representation, and their respective parameters were selected via evolutionary algorithms. The objective is to jointly explore the potentialities of both symbolic representation techniques and machine committee for biometric recognition based on ECG signals. Experimental results conducted on two publicly available datasets indicate that the proposed approach may improve recognition performance when compared to traditional techniques.

Keywords: Biometric Systems; Biomedical Signals; Electrocardiogram; Symbolic Representation; Machine Committees; Evolutionary Algorithms.

Lista de figuras

Lista de algoritmos

Lista de tabelas

Sumário

[1] De acordo com a Associação Brasileira de Normas Técnicas. NBR 6023.

1 Posicionamento e Motivação da Pesquisa

Na sociedade atual, a identificação automática, precisa e rápida dos indivíduos é uma necessidade cada vez mais crescente. Técnicas tradicionais de reconhecimento, tais como, números PIN (Personal Identification Number), senhas e cartões de identificação têm gerado preocupações pela facilidade de fralde, roubo e utilização por indivíduos não-autorizadas (JAIN; FLYNN; ROSS, 2007). De acordo com o relatório da CSN (Consumer Sentinel Network (REPORT, 2012)) de 2012, o roubo de identidade foi o número um na categoria de denúncia nos Estados Unidos (um milhão de casos de reclamações de consumidores). Como o roubo de identidade pode assumir diferentes formas, a falsificação de documentos/benefícios públicos (46%) foi a mais proeminente, seguida por fraude de cartão de crédito (13%), fraude usando telefones e similares (10%), fraudes bancárias (6%), fraudes relacionadas ao emprego (5%) e outras. Dentre esses casos, a falsificação de identidade aparenta ser o maior problema.

Nesse contexto, existe um crescente incentivo ao uso da biometria. Por exemplo, a autenticação multifator que combina a biometria com as técnicas convencionais (O'GORMAN, 2003) e até mesmo a substituição das técnicas tradicionais de reconhecimento. Biometria abarca um conjunto de métodos automatizados para verificar ou reconhecer a identidade de uma pessoa baseado em suas características físicas ou comportamentais (WAYMAN et al., 2005). No campo da biometria, a análise de características físicas e comportamentais dos seres humanos coletadas são empregues na identificação/verificação de indivíduos (AL-RAISI; AL-KHOURI, 2008), os sistemas biométricos são de grande importância e amplamente aplicados em automações para o reconhecimento de indivíduos. O termo modalidade biométrica se refere a um tipo específico de característica coletada. As principais modalidades biométricas são impressão digital (JAIN; FLYNN; ROSS, 2007), face (PHILLIPS et al., 2000), voz (GOMES, 2007), palma da mão (JAIN; FLYNN; ROSS, 2007), e íris (DAUGMAN, 2007; ZIAUDDIN; DAILEY, 2009). O processo de identificação biométrica pode ser dividido nas seguintes etapas: aquisição/segmentação, extração/seleção de características, comparação de características e armazenamento (JAIN; FLYNN; ROSS, 2007). A aquisição corresponde a coleta do que será a base para o reconhecimento biométrico como imagens, vídeos, sinais, impressões e etc, a segmentação identifica e delimita as regiões dos dados coletados subsequentemente. A extração de características tem como

objetivo extrair dos dados que foram segmentados as características fundamentais para o reconhecimento dos indivíduos. No armazenamento as características base para o reconhecimento são mantidas e a comparação tem o papel de estabelecer uma medida de distância entre as características de indivíduos distintos.

O primeiro trabalho sobre reconhecimento biométrico automático foi baseado na impressão digital publicado por Mitchell Trauring na revista Nature em 1963 (TRAURING, 1963). O reconhecimento biométrico automático com base em outras modalidades, tais como voz (PRUZANSKY, 1963), face (CHAN; BLEDSOE, 1965) e assinatura (MAUCERI, 1965) também surgiu na década de 1960. A seguir, os reconhecimentos biométricos baseados na geometria da mão (OTHERS, 1971), íris (DAUGMAN, 1993), dentre outras modalidades biométrica foram propostas. Wayman et al. (2005) traçaram os principais desenvolvimentos em biometria nos Estados Unidos de 1960 a 1990, e observou o seguinte: em uma breve visão geral da história que abrange os sistemas biométricos evidência que muito do que é pensado ser "novo" na biometria foi realmente apreciado nas décadas atrás. Também fica claro a necessidade do aprimoramento contínuo da biometria para a sociedade. Esta dissertação tem o objetivo de aprimorar a biometria, visto que tenta trazer para biometria novas formas de extração de características para o sinal de ECG, sendo o ensemble de representações de séries temporais uma forma de extrair características e classificar séries temporais.

Avanços significativos no reconhecimento biométrico foram alcançados, a importância da biometria como forma confiável para o reconhecimento de indivíduos foi estabelecida (JAIN; NANDAKUMAR; ROSS, 2016). No entanto, mesmo com os avanços nos sistemas biométricos e sua eficácia, pontos de melhorias foram levantados, tais como (ROSS; JAIN, 2003):

- *Coleta*: dados capturados pelos sensores são frequentemente afetados por ruído devido às condições do ambiente (insuficiência de luz, energia, etc) ou devido às condições fisiológicas e físicas do usuário (frio, dedos machucados, etc).
- *Distinguibilidade*: nem todas as modalidades biométricas tem o mesmo grau de distinguibilidade (por exemplo, sistemas biométricos baseados na geometria da mão são menos seletivos que aqueles baseados na impressão digital).
- *Variabilidade intraclasse*: significa que as modalidades biométricas podem variar com o tempo para uma mesma pessoa e essa variação é imprevisível de pessoa para

pessoa. Por exemplo, à medida que uma pessoa envelhece pode ocorrer mudanças na sua voz.

Além das questões mencionadas anteriormente, existem fraudes cometidas contra os sistemas biométricos, isso porque nenhum sistema é totalmente imune a ataques, porém a eficiência comprovada dos sistemas biométricos a décadas e as novas propostas que mitigam os pontos falhos fazem com que os sistemas biométricos sejam uma realidade na atualidade. Galbally-Herrero et al. (2006) recorreram a massa de plastimodelismo e silicone para gerar uma impressão digital falsa, e que foi aplicada para burlar dois sistemas biométricos por impressão digital em um ataque direto com algum sucesso. Porém para contrapor as fraudes existem melhorias nos sistemas biométricos, Marcialis, Roli e Tidu (2010) propuseram a análise dos poros da pele na impressão digital para reconhecer características de vivacidade em uma coleta. Há casos também em que as características biométricas são intencionalmente removidas para evitar a verdadeira identidade (ALLEN, 2009).

Recentemente, pesquisas têm sido realizadas para a geração de novas modalidades biométricas. Características internas ao organismo humano têm sido investigadas, como padrões das veias (ZHOU; KUMAR, 2011), odor (PANKANTI; BOLLE; JAIN, 2000), biometria cognitiva (HOFFMANN et al., 2008; REVETT; DERAVI; SIRLANTZIS, 2010), dentre outras biometrias (ADEOYE, 2010). Da mesma forma, os sinais biomédicos constituem outra categoria de novas modalidades de reconhecimento biométrico, que engloba sinais que são tipicamente utilizados em diagnósticos clínicos. Sinais biomédicos, especialmente eletrocardiograma (ECG), eletroencefalograma (EEG) e eletromiograma (EMG), são considerados modalidades biométricas emergentes (ISRAEL et al., 2005; SUN, 2008; PALANIAPPAN; MANDIC, 2007; SINGH; GUPTA, 2011).

Novas características são aplicadas no reconhecimento, a coleta das características e o tempo no reconhecimento são fatores importantes para esses sistemas biométricos. Esses requisitos influenciaram o desenvolvimento de novos sensores principalmente sensores a seco, que são fáceis de serem fixados mesmo por pessoas não treinadas, o campo da biometria biomédica floresceu. Além disso, o rápido avanço entre 2001 e 2010 foi apoiado pelo fato de que o processamento dos sinais biomédicos (ou biossinal) já tinha alcançado grandes progressos para fins de diagnóstico (HAMILTON; TOMPKINS, 1986; THAKOR; ZHU, 1991). Os sinais biomédicos podem ser combinados a outras modalidades biomédicas para

o aperfeiçoamento no reconhecimento (ISRAEL et al., 2003; FATEMIAN; AGRAFIOTI; HATZINAKOS, 2010). Os sinais biomédicos requerem a adoção de processamentos de sinais específicos e representações destinadas à extração de informações relevantes nesses sinais (KULAHCIOGLU; OZDEMIR; KUMOVA, 2008; HAYKIN et al., 2009; FU, 2011).

Reconhecimento de identidade baseado em sinais de ECG remonta ao trabalho pioneiro de Biel et al. (2001), Irvine et al. (2001), Kyoso e Uchiyama (2001). A premissa desses e outros trabalhos é que o ECG contém informações relativas à operação elétrica do coração, essas atividades são potencialmente capazes de identificar indivíduos (SIMON; ESWARAN, 1997). O sinal de ECG é altamente individualizado, na medida em que depende de propriedades funcionais e estruturais incluindo condutividade do coração e outros tecidos. A principal hipótese compartilhada por esses estudos é que a atividade elétrica do coração, como coleta do ECG, está relacionada às características individuais, logo pode ser utilizado em sistemas de reconhecimento de identidade de alto desempenho (ODINAKA et al., 2012). O sinal de ECG pode ser descrito por uma série temporal.

Uma série temporal é uma coleção de observação realizada cronologicamente, a qual pode ser obtida de medidas de forma geral, como flutuações diárias de ações do mercado, ECG, EEG, observações experimentais médicas, biológicas e etc. Há vários tipos de pesquisas relacionadas a séries temporais, como indexação (SHIEH; KEOGH, 2009; CAMERRA et al., 2010), classificação (GEURTS, 2001; XI et al., 2006) e representação de séries temporais (LKHAGVA; SUZUKI; KAWAGOE, 2006b; LIN et al., 2007), que são pesquisas relacionadas a área de mineração de dados em séries temporais (ESLING; AGON, 2012). No contexto da mineração de dados em séries temporais, em especifico o reconhecimento de padrões, as representações de séries temporais têm influência no sucesso da classificação, essas representações vêm sendo aplicadas na extração de características alcançando bons resultados (LIN et al., 2007). As representações de séries temporais podem apresentar resultados melhores do que a série temporal bruta (LIN et al., 2007).

Atualmente, vários trabalhos de representação de séries temporais foram propostos (DING et al., 2008), tais como, Aproximação Agregada por Parte (do inglês Piecewise Aggregate Approximation - PAA) (KEOGH et al., 2001a), Aproximação Constante por Parte Adaptativa (do inglês Adaptive Piecewise Constant Approximantion - APCA) (KEOGH et al., 2001b), Aproximação Agregada Simbólica (do inglês Symbolic Aggregate approXimation - SAX) (LIN et al., 2007), Aproximação Linear Por Parte Indexável (do inglês Indexable Piecewise Linear Approximation - PLA) (CHEN et al., 2007), Aproximação

Por Segmento Derivativo de Séries Temporais (do inglês Derivative time series Segment Approximation - DSA) (GULLO et al., 2009), Pontos Importantes Perceptualmente (do inglês Perceptually Important Points - PIP) (FU et al., 2008), Transformada Discreta de Wavelet (do inglês Discrete Wavelet Transform - DWT) (CHAN; FU, 1999), Polinômio de Chebyschev (do inglês Chebyschev Polynomials - CHEBY) (CAI; NG, 2004), Transformada Discreta de Fourier (do inglês Discrete Fourier Transforma - DFT), Decomposição de Valores Singulares (do inglês Single Value Decomposition - SVD) (FALOUTSOS; RANGANATHAN; MANOLOPOULOS, 1994) e Transformada Discreta de Cosseno (do inglês Discrete Cosine Transform - DCT) (KORN; JAGADISH; FALOUTSOS, 1997), dentre outras representações de séries temporais.

Aproximação por parte representa as séries temporais com funções descontinuas que sofrem de alguns problemas: 1) a aproximação por parte pode conduzir a erros e desvios após a redução da dimensionalidade em relação a série temporal original, mas esses fatores não caracterizam uma ineficiência dessas representações, pois alguns trabalhos já provaram sua relevância (LIN et al., 2007); 2) em algumas representações de séries temporais não existe uma relação direta entre a série temporal original e sua representação.

Aproximação continua minimiza o desvio máximo (aproximação minmax) entre a série temporal e as funções continuas que são obtidas com polinômios, como nos polinômios de Chebyshev. Aproximação polinomial usualmente sofre do fenômeno de oscilação, especialmente para polinômios de alto grau.

Sendo assim, há inúmeras representações de séries temporais simbólicas, entretanto muitas representações são paramétricas e não existem um consenso na busca por esses parâmetros. Por outro lado, algumas representações isoladamente podem produzir resultados ruins, mas quando combinadas com outras representações podem melhorar o desempenho alcançado por uma representação. Mediante esses fatores a busca por parâmetros e a combinação de representações de séries temporais em um ensemble de técnicas de representação simbólica são bastante promissores. Basicamente, o ensemble de técnicas de representação simbólica é inspirado na ideia do comitê de máquina de aprendizado, nesse trabalho mais especificamente a composição (do inglês ensemble). As máquinas de aprendizado proporcionam individualmente resultados satisfatórios e diversos, sendo que possivelmente um aumento da acurácia na composição de classificadores pode ser alcançado. A ideia de comitês de máquinas data de 1965, quando Nilsson (1965) considerou a estrutura de uma rede neural composta de uma camada de perceptrons elementares seguidos por um

perceptron responsável por realizar um esquema de votação na segunda camada. Comitês de máquinas são aproximadores universais (HAYKIN et al., 2009) e podem se apresentar em versões estáticas, denominadas composição de componentes (do inglês ensemble), ou então em versões dinâmicas, denominadas misturas de especialistas. Nesse trabalho, a combinação de representações de séries temporais simbólicas será adotada para o sinal ECG visando o reconhecimento biométrico.

Zhou, Wu e Tang (2002) propuseram um comitê de máquinas de aprendizado composto por redes neurais, um dos problemas enfrentado é a seleção de redes neurais e pesos para a composição do comitê, isso porque a combinação de soluções possíveis é grande, diante dessas circunstâncias foram adotados algoritmos evolutivos para a busca dos componentes do comitê. No ensemble de técnicas de representação simbólica a grande quantidade de combinações de representações e seus parâmetros pode tornar inviável a simulação de todas as possíveis soluções, para selecionar a melhor solução para o ensemble essa dissertação adota algoritmos evolutivos.

1.1 Objetivo da pesquisa

O objetivo desta dissertação é avaliar técnicas de representações de séries temporais simbólicas e o ensemble dessas para a tarefa de classificação visando reconhecimento biométrico. A abordagem está restrita ao reconhecimento biométrico baseado em sinais de ECG. Algoritmos evolutivos foram utilizados para seleção dos parâmetros para o ensemble de técnicas de representação simbólica.

Para alcançar este objetivo geral, o projeto possuiu ainda os seguintes objetivos específicos:

- Buscar por técnicas de representação simbólica para o uso no ensemble;
- Implementar um ensemble de técnicas de representação simbólica;
- Efetuar as simulações para o reconhecimento de padrões biométricos;
- Avaliar o desempenho da aplicação ensemble de técnicas de representação simbólica para o reconhecimento biométrico baseado em sinais de ECG.

1.2 Organização do texto

O capítulo 1 descreve a motivação, os objetivos e a organização do texto. O capítulo 2 apresenta um breve panorama sobre sistemas biométricos. O capítulo 3 descreve as principais representações de séries temporais aplicadas a abordagem proposta. O capítulo 4 aborda uma visão abrangente dos algoritmos evolutivos e classificadores, que fazem parte da abordagem proposta. O capitulo 5 descreve e ilustra a abordagem proposta. O capitulo 6 apresenta os resultados obtidos com a abordagem proposta. Já o capítulo 6 é contemplado por comentários e discussões sobre os resultados alcançados.

2 Sistemas Biométricos

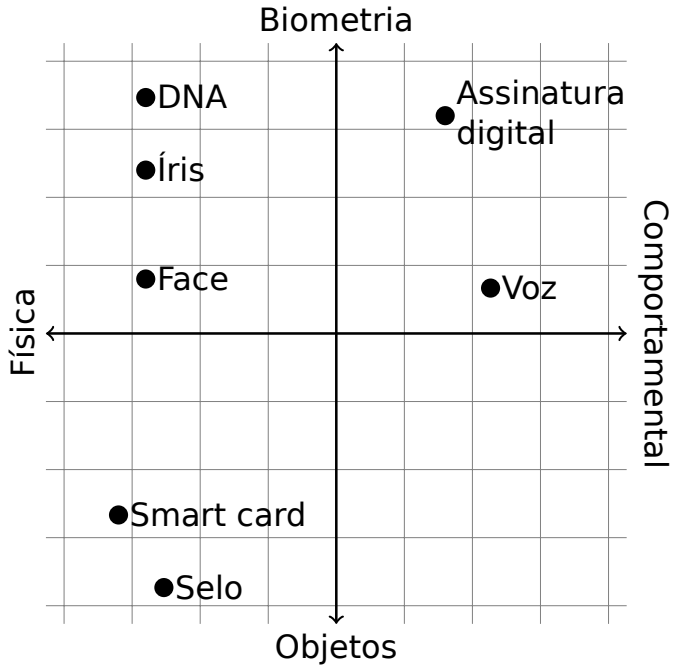
A necessidade de reconhecer indivíduos vai além de ambientes seguros, existem diversas aplicações a biometria, desde uma assinatura em um contrato ao reconhecimento de uma individuo pelo DNA (do inglês deoxyribonucleic acid - DNA; do português ácido desoxirribonucleico), isso é possível devido as mais diversas propostas existentes na biometria. Cada modalidade biométrica (impressão digital, íris, face e etc) tem especificidade que em termos de aplicação, segurança, dentre outros aspectos, esse é um dos motivos que as melhorias propostas e as novas modalidades biométricas sejam tão importantes. Diversos estudos sobre as modalidades biométricas que adotaram características físicas e comportamentais dos indivíduos, porém ultimamente as modalidades biométricas são propostas como modalidades biométricas baseadas nos sinais de ECG e EEG.

Biometria é o estudo de características físicas e comportamentais dos seres vivos (MALTONI et al., 2009) visando sua identificação. A modalidade biométrica é a categorização atribuída a um sistema biométrico, que depende do tipo de característica física ou comportamental adotada como entrada do sistema. Dentre as diversas modalidades biométricas destacam-se impressão digital, face, íris, voz, geometria da mão, dentre outras. O processo de identificação através da biometria não deve ser facilmente burlado, esquecido ou compartilhado. A biometria é relativamente mais confiável do que os processos de reconhecimento de pessoas baseado em objetos de autenticação, como chaves (por exemplo, chave, cartão de identificação) ou conhecimento (por exemplo, senha ou PIN).

A figura 1 ilustra o contraste da biometria com as técnicas convencionais e da biometria baseada em características físicas com a biometria baseada em características comportamentais. Qualquer característica física ou comportamental pode ser utilizada como modalidade biométrica, desde que satisfaçam as seguintes condições (JAIN; ROSS; PRABHAKAR, 2004):

- Universalidade: as características devem estar presentes em todas as pessoas;
- Unicidade: as características devem ser suficientemente distintas entre as pessoas;
- Permanência: as características devem ser invariantes por um período de tempo;
- Coletável: as características devem ser passíveis de serem mensuradas quantitativamente.

Figura 1 – Identificação e/ou validação pessoas



Fonte: Henrique Passos, 2018

Para que uma modalidade biométrica seja empregada na prática ela deve atender os seguintes quesitos (JAIN; ROSS; PRABHAKAR, 2004):

- Desempenho: o sistema deve ser preciso e rápido para que a resposta correta seja apresentada em um espaço de tempo aceitável;
- Aceitabilidade: o sistema deve ser aceito pelas pessoas que estão interagindo com ele;
- Circunvenção: representa o quão um sistema é robusto a fraudes.

2.1 Sinais biomédicos aplicados a biometria

Novas modalidades biométricas têm sido propostas na literatura, essas modalidades trouxeram novas perspectivas para a biometria, novas oportunidades para o reconhecimento de indivíduos mais robustos a fraudes (LUZ; MENOTTI; SCHWARTZ, 2014). Órgãos como o coração, cérebro e músculo são compostos de tecidos proteicos chamados células do miocárdio, glias e fibras musculares respectivamente. Portanto, os sinais elétricos evocados a partir desses órgãos devem mostrar a singularidade dos indivíduos (SIMON; ESWARAN, 1997). Existem fortes evidências da dificuldade da falsificação biométrica dos batimentos cardíacos humano (PLATANIOTIS; HATZINAKOS; LEE, 2006).

2.1.1 Eletrocardiograma

O eletrocardiograma (ECG) é o registro da atividade elétrica durante o ciclo cardíaco por meio de um aparelho denominado eletrocardiógrafo, sendo efetivo, simples e não invasivo, o qual fornece informações sobre o funcionamento do coração. A análise dos sinais de ECG promove informações valiosas sobre o estado cardíaco de uma determinada pessoa (SAHOO; BEHERA; ARI, 2011). Muitos métodos baseados em computação vêm sendo propostos para automatizar o diagnóstico de patologias. O princípio fundamental de tais métodos é o uso de técnicas de reconhecimento de padrões.

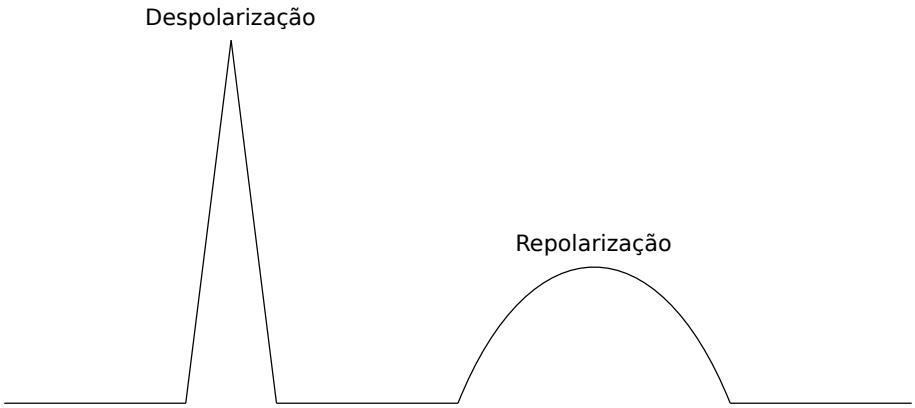
Os sinais de ECG são caracterizados por cinco formas de ondas importantes P, Q, R, S e T. Normalmente, as ondas Q, R, S são agrupadas para formar o que é chamado de complexo QRS. A onda P representa a duração do tempo de despolarização dos átrios, o complexo QRS a despolarização ventricular e subsequentemente a repolarização ventricular com a onda T. A forma de onda da repolarização atrial é obscurecida pela contração ventricular, uma vez que o ventrículo tem maior massa muscular. Outras propriedades importantes dos sinais de ECG são os intervalos entre as ondas. Os intervalos mais importantes são PQ, intervalo QRS, QT, ST e RR. O intervalo RR é o tempo de duração entre a onda R de dois batimentos cardíacos consecutivos. O desempenho de grande parte dos sistemas de diagnósticos para o coração, que utilizam características extraídas do sinal de ECG, dependem da detecção das cinco formas de ondas mencionadas acima.

Quando um tecido muscular estriado cardíaco apresenta no seu interior um potencial elétrico negativo temos uma musculatura cardíaca em repouso, também conhecida como polarizada, mas quando o processo de despolarização ocorre o seu interior torna-se um potencial elétrico positivo, o que resulta na contração das fibras musculares chamadas de miócitos (DUBIN, 1989). A figura 2 exemplifica as fases de despolarização e repolarização que é apresentado no eletrocardiograma.

O resultado da medição do eletrocardiógrafo é uma série temporal que representa o sinal de ECG, o qual contem padrões resultantes das atividades elétricas. A figura 3 mostra as ondas P, Q, R, S, T e U.

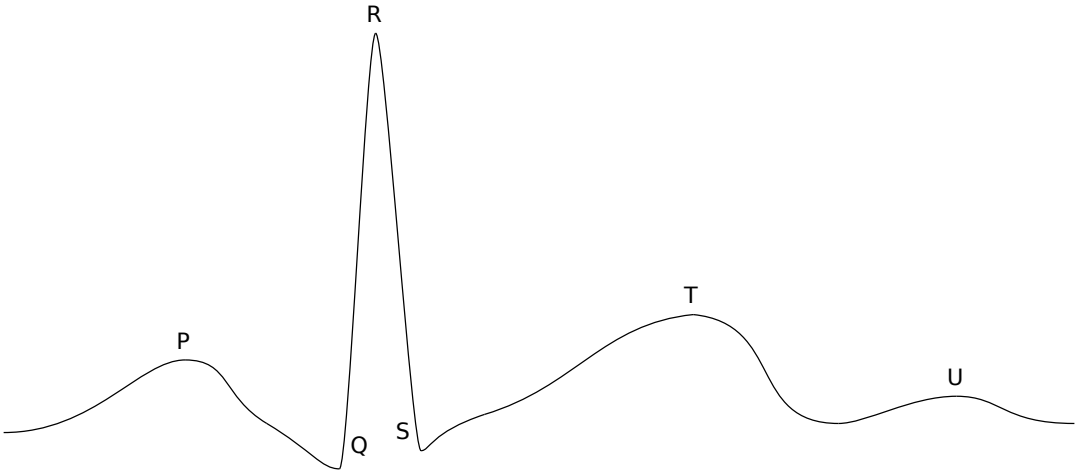
Características gerais, tais como sexo, peso corporal, configuração peitoral e fatores diversos provocam alterações dos batimentos cardíacos, que agem no ECG gerando padrões

Figura 2 – Polarização



Fonte: Henrique Passos, 2018

Figura 3 – Sinal ECG



Fonte: Henrique Passos, 2018

característico. Logo, esses padrões podem ser como base para identifica-los (SIMON; ESWARAN, 1997).

2.1.2 Aplicação de sinais de eletrocardiograma em biometria

Biel et al. (2001) foram pioneiros no reconhecimento biométrico baseado em ECG, seus experimentos consistem na coleta do ECG por 12 canais para cada pessoa e na extração de 30 características fiduciais para cada batimento cardíaco. As características fiduciais estão relacionadas com distâncias entre ondas, ângulos de picos/vales, áreas e etc, como ilustrado na figura 4, nessa figura QS, PS, QT, PQ, ST e PT são características fiduciais que representam distâncias entre ondas do sinal de ECG, o QS é a característica baseada na distância entre as ondas Q e S, por outro lado P, QRS e T são características

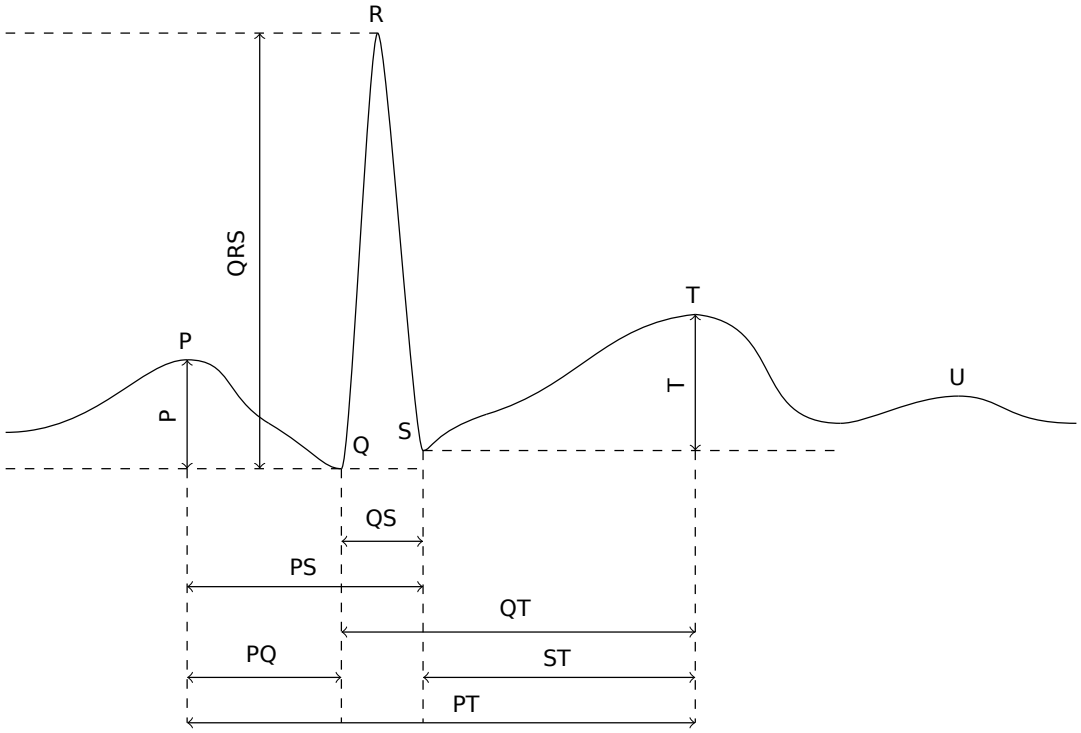
fiduciais que mantem a altura de ondas. A Modelagem Flexível Independente de Classes Análogas (do inglês Soft Independent Modeling of Class Analogy - SIMCA) (ESBENSEN; SCHÖNKOPF; MIDTGAARD, 1994) foi adotada para classificar os indivíduos.

Características não fiduciais foram empregadas por Chan et al. (2008) na identificação de indivíduos, como Coeficiente de Correlação (do inglês Correlation Coefficient - CCORR) e Medida de Distância Wavelet (do inglês Wavelet Distance Measure - WDIST). As características não fiduciais têm relação com o ECG, mas não diretamente com os pontos fiduciais do ECG. Com isso não existe a necessidade da detecção de todas as ondas de um ECG, em alguns casos pode existir a necessidade da identificação dos batimentos cardíacos para exercer a extração de características não fiduciais batimento a batimento.

Em relação a coleta, o eletrocardiógrafo é amplamente empregado nesse tipo de modalidade biométrica, em alguns casos esse equipamento pode ser grande em termos de dimensões, porém existem alternativas portáteis e financeiramente de baixo custo. Lourenço, Silva e Fred (2011) obtiveram 94,3% de acento no reconhecimento biométrico de 16 pessoas utilizando um equipamento de tamanho reduzido que coleta o ECG através dos dedos de pessoas. No mesmo sentido, Shen, Tompkins e Hu (2010) apresentaram um equipamento que coleta o ECG da palma das mãos de um indivíduo, mesmo o sinal coletado nas mãos apresentar maior quantidade de interferências com relação ao sinal coletado do peito, os resultados da abordagem proposta chegaram a 98% de acurácia. Rede de área corporal sem fio (do inglês Wireless Body Area Network - WBAN) é um equipamento desenvolvido que opera de forma autônoma na coleta de dados de diversos sensores médicos, mesmo sendo um equipamento emergente é largamente estudada para diversas aplicações, Wang et al. (2011) apresentaram uma solução de segurança que aplicou esse equipamento para coletar sinais de ECG para o reconhecimento biométrico. Chun et al. (2016) apresentaram a utilização de uma pulseira responsável pelo reconhecimento biométrico baseado em ECG, com apenas o toque de um dedo no dispositivo os batimentos cardíacos são coletados e o indivíduo é identificado.

Um sistema biométrico (JAIN; ROSS; PRABHAKAR, 2004) baseado em ECG é descrito no diagrama de bloco mostrado na figura 5. O ECG é coletado, convertido para um formato digital, pré-processado e pôr fim a extração das características é efetuada. As características são utilizadas na classificação dos indivíduos visando o reconhecimento (modo identificação) ou autenticação (modo verificação) biométrica do indivíduo (CANENTO et

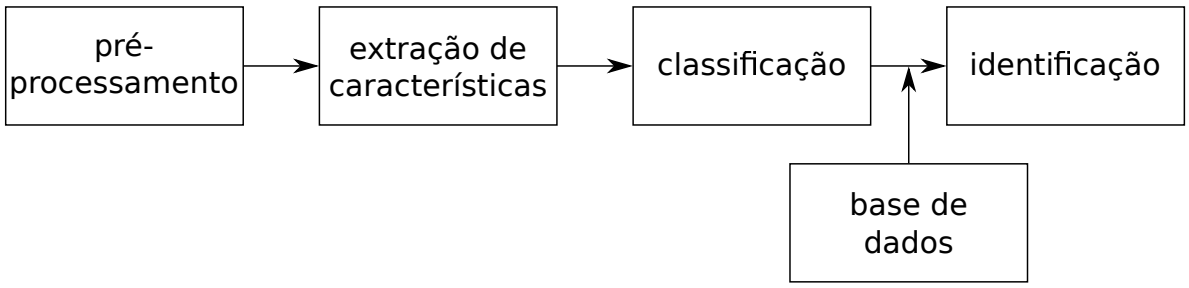
Figura 4 – Sinal ECG extração de características



Fonte: Henrique Passos, 2018

al., 2013; LOURENÇO; SILVA; FRED, 2012). Odinaka et al. (2012) estabeleceram uma análise comparativa entre as abordagens propostas sobre a biometria baseada em ECG.

Figura 5 – Sistema biométrico baseado em sinais de ECG



Fonte: Henrique Passos, 2018

2.2 Sistemas Biométricos baseado em Sinais de ECG

2.2.1 Detecção do complexo QRS

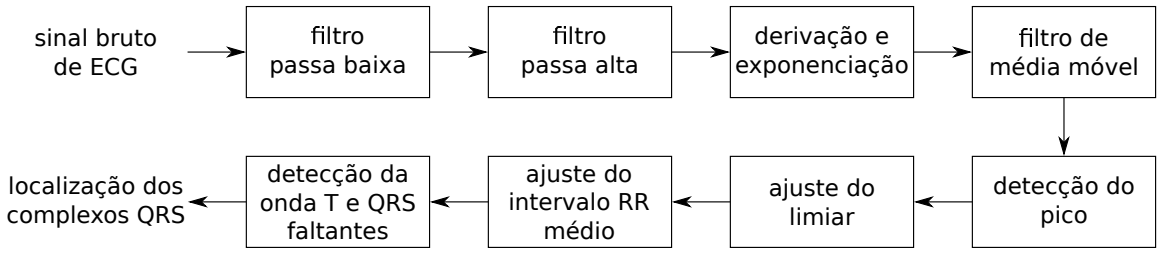
Para a segmentação dos batimentos, são utilizadas técnicas de processamento digital de sinais com o objetivo de detectar os complexos QRS, as transformadas wavelets (ADDISON, 2005) são amplamente utilizadas em sinais biomédicos; e, mais recentemente,

o método da operação diferencial (do inglês Difference Operation Method - DOM) (YEH; WANG, 2008), dentre outras.

2.2.2 Algoritmo de Pan-Tompkins

Nesta dissertação foi utilizado o algoritmo de Pan-Tompkins (PAN; TOMPKINS, 1985), o qual tem como propósito identificar o complexo QRS. A figura 6 descreve todo o processo de detecção. Inicialmente, com o objetivo de atenuar possíveis ruídos os filtros de passa baixa e passa alta são aplicados ao sinal, esses dois filtros em cascata são equivalentes ao filtro passa banda com largura de banda de 5 a 11 Hz. Informação sobre o angulo do complexo QRS é enfatizado no filtro derivativo, na sequência o sinal é elevado ao quadrado reforçando esse angulo. O filtro de média móvel calcula a média das últimas 30 amostras do sinal, com essa integração um pico é gerado no sinal representando o angulo, largura e localização do QRS. Posteriormente, um limiar é definido com a média desse sinal e todos os valores maiores do que esse limiar corresponde a um QRS. Os complexos QRS localizados são base para a busca dos possíveis QRS faltantes, o algoritmo de Pan-Tompkins define o intervalo RR médio e uma nova busca é realizada nesses intervalos para encontrar os QRS faltantes.

Figura 6 – Algoritmo de Pan-Tompkins



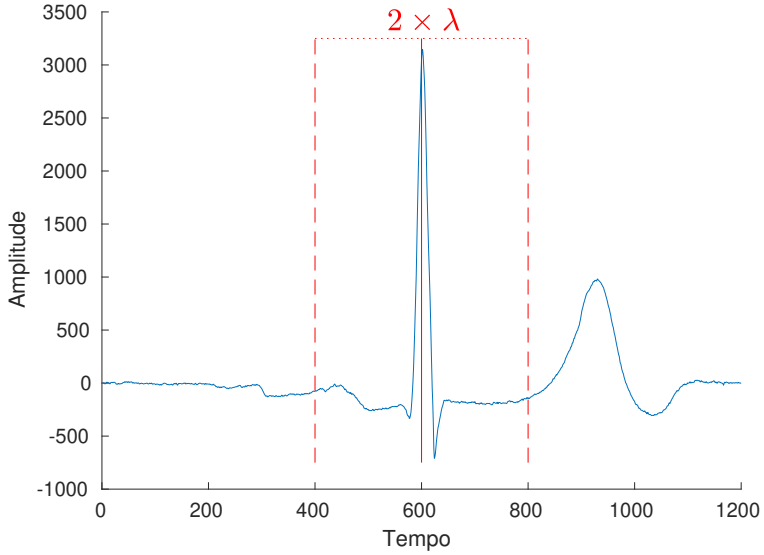
Fonte: Henrique Passos, 2018

2.2.3 Pré-processamento do ECG

O que caracteriza o ECG são as atividades do coração, o complexo QRS é a região de maior atividade do coração, sendo aproximadamente o centro do batimento cardíaco o ponto mais característico dentro do ECG (LUGOVAYA, 2005; WÜBBELER et al., 2007; CHIU; CHUANG; HSU, 2008), por esse motivo a busca do complexo QRS mesmo na extração de características não-fiduciais é tão importante. O algoritmo de Pan-Tompkins (PAN;

TOMPKINS, 1985) tem o papel fundamental na identificação dos batimentos cardíacos, mais precisamente o pico R centro do complexo QRS alvo da extração de características não-fiduciais da abordagem proposta. Com a identificação de cada batimento uma janela valores ao redor do pico R é extraída, sendo *janela do QRS* valores a direita e *janela do QRS* valores a direita, como ilustra a figura 7.

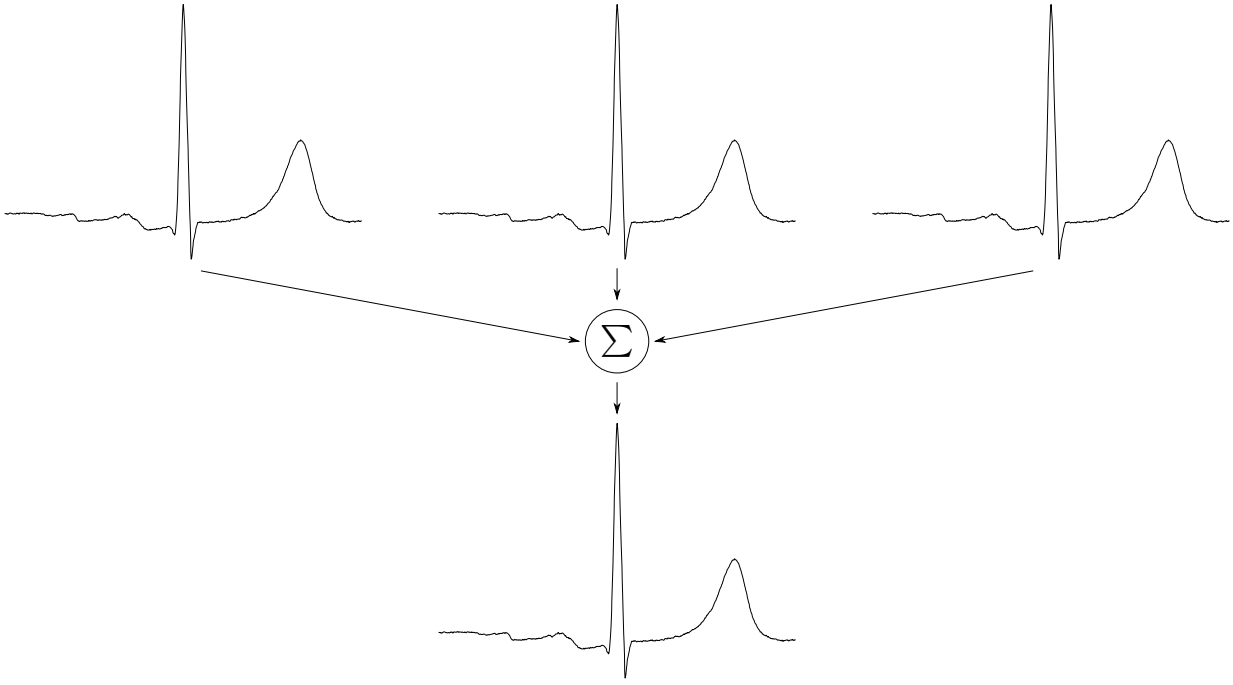
Figura 7 – Região característica de um batimento



Fonte: Henrique Passos, 2018

Após a extração dos batimentos cardíacos possíveis interferências podem estar representes nos batimentos. Com o objetivo de minimizar essas possíveis interferências o sinal médio de ECG (do inglês Signal-Averaged ECG - SAECG) (CHAN et al., 2008) é aplicado aos batimentos cardíacos segmentados. O sinal médio de ECG realiza uma suavização do sinal através da média entre batimentos consecutivos (em relação a sua extração) gerando um batimento médio, o parâmetro *janela média* é a definição da quantidade de batimentos consecutivos que serão utilizados para a geração de um batimento médio. O sinal médio de ECG é ilustrado na figura 8, nessa a suavização é feita através de três batimentos (*janela média*) para a geração de um batimento médio.

Figura 8 – Sinal médio de ECG



Fonte: Henrique Passos, 2018

3 Representações de Séries Temporais

Desde o início da humanidade o homem tem a necessidade de contar ou estabelecer medidas das coisas ao seu redor, essas coisas mudam com o passar do tempo. Essa necessidade ainda se reflete atualmente de forma sofisticada em gráficos com informações sobre ações na bolsa de valores, temperatura em uma determinada localidade, dentre outros exemplos. Uma série temporal é um tipo de dado que mantém os valores coletados e armazenados de forma que se possa saber a sua cronologia. A análise desse tipo de dado é chamada de mineração de dados em séries temporais. Dentro da mineração de dados em séries temporais existem diversas áreas que tem como foco investigar formas eficiente de representar dados dessa natureza. Nesse capítulo, será apresentado diversas formas de representar simbolicamente uma série temporal.

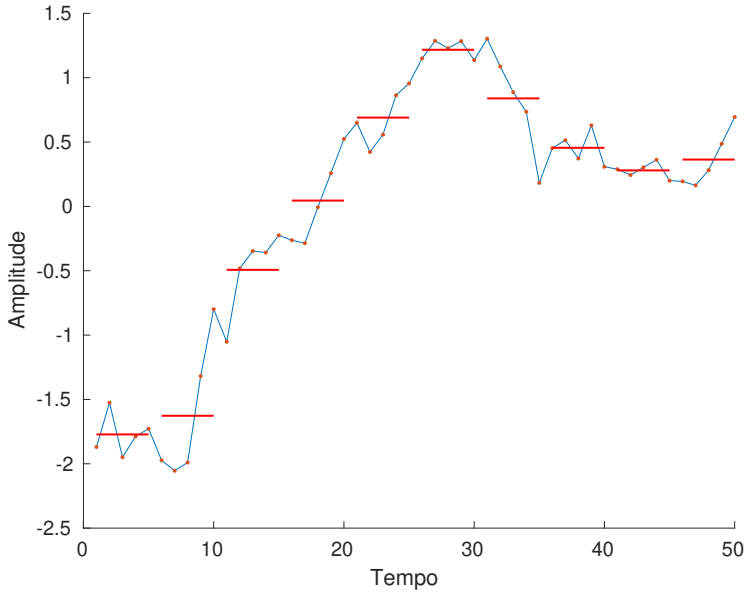
Mineração de dados em séries temporais baseia-se na extração de informações não triviais, tais como, padrões, tendências e relacionamentos nos dados. Inúmeros trabalhos envolvendo técnicas de aprendizado de máquina e estatística para indexação, classificação, clusterização e representação de séries temporais têm sido propostos (KEOGH; KASETTY, 2003; LAXMAN; SASTRY, 2006). Nesse capitulo, é descrito diversas formas de representações simbólicas para séries temporais.

Nas figuras que ilustram as representações de séries temporais, a coordenada horizontal representa uma unidade de tempo e a coordenada vertical uma unidade de medida qualquer. Pelo fato das representações não terém uma relação direta entre a série temporal original e sua representação, as unidades nas ilustrações não foram especificadas.

3.1 Aproximação Agregada por Parte

Aproximação Agregada por Parte (do inglês Piecewise Aggregate Approximation - PAA) (KEOGH et al., 2001a), foi proposta por Keogh et al. (2001a). O PAA realiza a redução da dimensionalidade das séries temporais empregando a aproximação por partes. Essa aproximação realiza a divisão da série temporal em partes consecutivas de tamanho igual chamadas de segmentos e usa o valor médio de cada segmento para representá-lo, como ilustrado na figura 9.

Figura 9 – Representação do PAA



Fonte: Henrique Passos, 2018

O parâmetro (ν) define o tamanho dos segmentos, esses segmentos representam as regiões que serão agregadas na redução da dimensionalidade. Todas as representações de séries temporais baseadas no PAA necessitam que este parâmetro seja definido. A equação 1 representa a redução de dimensionalidade realizada pelo PAA.

Dado uma série temporal $X = X_1, X_2, \ldots, X_N$ de tamanho N, o PAA calcula o valor médio de cada segmento com o objetivo de reduzir a série temporal X de tamanho N para n resultando na série temporal dada por $\hat{X} = \hat{X}_1, \hat{X}_2, \ldots, \hat{X}_n$, o parâmetro $\nu = \frac{N}{n}$.

$$\hat{X}_i = \frac{N}{n} \sum_{j=\frac{n}{N}(i-1)+1}^{\frac{n}{N}i} X_j \tag{1}$$

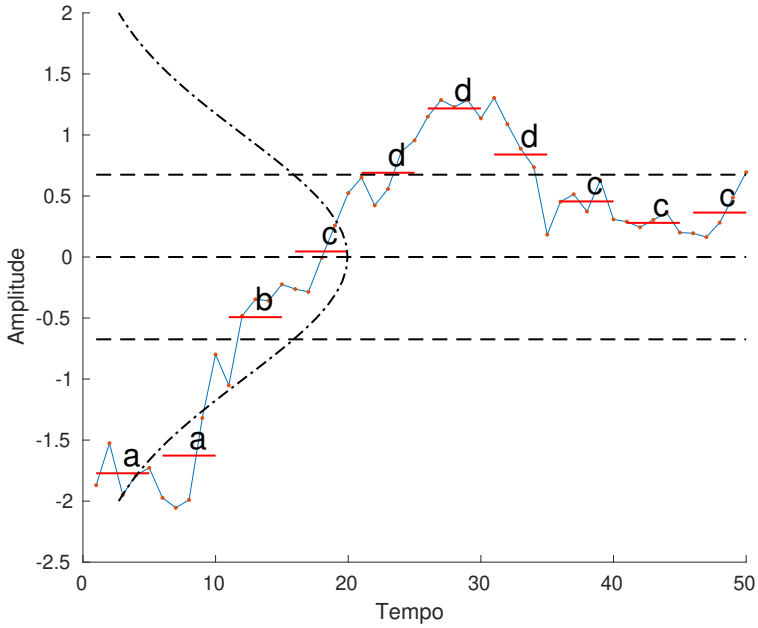
Lin et al. (2007) ressaltaram o quão intuitivo e simples é o PAA. Apesar da sua simplicidade é bastante promissor na redução de dimensionalidade em comparação com a Transformada de Fourier e Wavelet (KEOGH et al., 2001b; KEOGH; KASETTY, 2003; YI; FALOUTSOS, 2000). Devido a sua simplicidade, o PAA é adotado como base para diversas outras representações de séries temporais.

Em mineração de dados em séries temporais, a definição de uma medida de distância entre as séries temporais é de extrema importância. O PAA ajuda em estabelecer essa similaridade entre as séries temporais, visto que o PAA normaliza as séries temporais com média zero e desvio padrão igual a um, removendo qualquer erro indesejado decorrente de deslocamentos e amplitudes nas séries temporais.

3.2 Aproximação Agregada Simbólica

Aproximação Agregada Simbólica (do inglês Symbolic Aggregate approXimation - SAX) foi proposta por (LIN et al., 2003). Essa representação de séries temporais emprega o PAA para a redução da dimensionalidade. Inicialmente, o SAX realiza a redução de dimensionalidade por meio do PAA, em seguida é efetuado a simbolização dos valores médios resultantes da redução de dimensionalidade. Durante a simbolização, a proximidade da distribuição do conjunto de dados com a distribuição gaussiana proporciona a divisão da coordenada vertical por áreas equiprováveis, as divisões entre essas áreas são chamadas de pontos de quebra (do inglês breakpoints - β), como é ilustrada na figura 10.

Figura 10 – Representação do SAX



Fonte: Henrique Passos, 2018

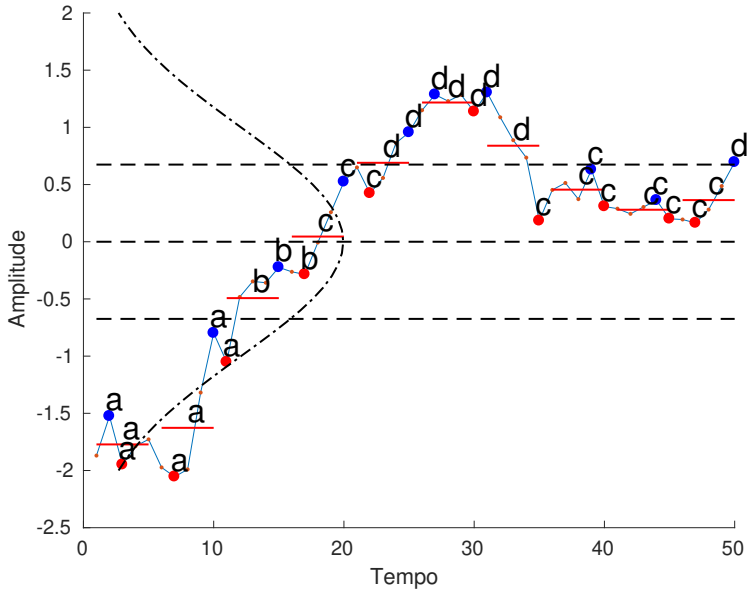
Estudos efetuados por Lin et al. (2003) com subsequências de tamanho 128 em 8 conjuntos de dados de séries temporais sugerem que subsequências de séries temporais são altamente gaussianas, esse foi um dos motivos da adoção da distribuição gaussiana pelo SAX. Sendo assim, os símbolos produzidos serão equiprováveis, mas apenas se a séries temporais normalizadas forem altamente gaussianas. Os parâmetros (ν) e (α) presentes no SAX representam o número de segmentos e o tamanho do alfabeto usado na simbolização respectivamente. Por exemplo, quando é definido o tamanho do alfabeto igual a 10 a série temporal será representada pelos símbolos $\{a, b, c, d, e, f, g, h, i, j\}$.

Passos et al. (2017) apresentaram um estudo referente aos parâmetros do SAX. Esse estudo investigou a importância de cada parâmetro para a classificação e a faixa de valores que maximizam o desempenho da classificação. As conclusões obtidas foram as seguintes: i) quanto menor o tamanho dos segmentos maior o desempenho, isso resulta em uma baixa redução da dimensionalidade da série temporal; ii) quanto maior o tamanho do alfabeto maior o desempenho, quanto maior a quantidade de símbolos distintos maior a capacidade de classificação, porém essa relação não é uma regra. Essa relação também pode mudar de acordo com a distribuição dos dados, sendo necessário muitas vezes estabelecer uma busca por esses parâmetros.

3.3 Aproximação Agregada Simbólica Estendida

Aproximação Agregada Simbólica Estendida (do inglês Extended Symbolic Aggregate approXimation - ESAX) (LKHAGVA; SUZUKI; KAWAGOE, 2006b; LKHAGVA; SUZUKI; KAWAGOE, 2006a), é uma variação do SAX, que além do valor médio da agregação, os valores de mínimo e máximo são igualmente simbolizados para cada segmento, como ilustrado na figura 11. O ESAX mantem os valores mínimo e máximo para cada segmento, por esse fato foi inicialmente idealizado para representação de séries temporais financeiras, na qual a ausência desses padrões motiva na grande probabilidade da perda de características importantes.

Figura 11 – Representação do ESAX



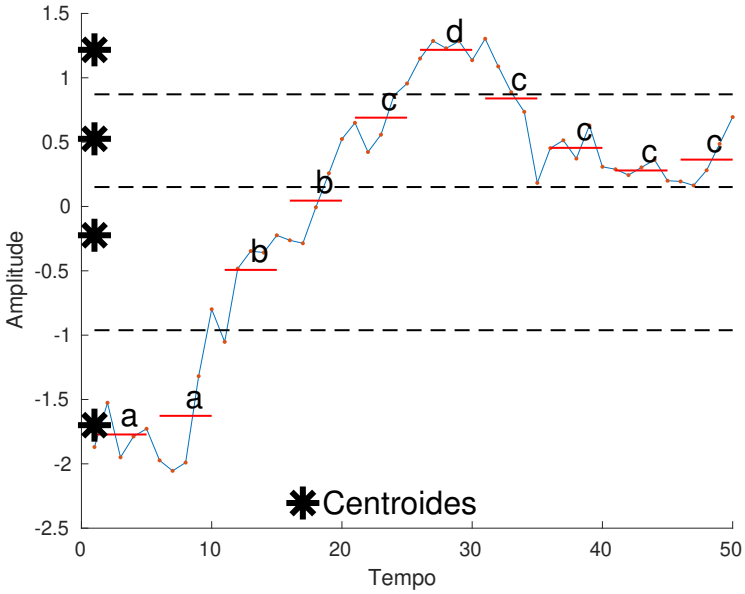
Fonte: Henrique Passos, 2018

Lkhagva, Suzuki e Kawagoe (2006b) apresentaram experimentos que comparam o desempenho do SAX e do ESAX, com o objetivo da busca por meio da força bruta de subsequência extraídas do próprio conjunto de dados alvo da busca. Os resultados apontam que o ESAX gera quase três vezes menos resultados falsos do que o SAX, isso mostra que o ESAX consegue gerar representações mais significativas que o SAX.

3.4 Aproximação Agregada Simbólica Adaptativa

As representações de séries temporais que assumem gaussianidade na coordenada vertical, consequentemente regiões equiprováveis, podem gerar símbolos de baixa representatividade nas regiões onde a série temporal não são altamente gaussianas. Aproximação Agregada Simbólica Adaptativa (do inglês Adaptive Symbolic Aggregate approXimation - ASAX) (PHAM; LE; DANG, 2010) realiza uma modificação na simbolização dos dados em relação ao SAX. O ASAX emprega técnica de agrupamento para definir as regiões a serem simbolizadas de forma adaptativa. Essa adaptabilidade possibilita representar séries temporais com diversas distribuições. Neste caso, após a redução de dimensionalidade, os dados são agrupados de acordo com uma técnica de agrupamento e cada grupo representa uma região a ser simbolizada, como ilustrado na figura 12.

Figura 12 – Representação do ASAX



Fonte: Henrique Passos, 2018

No ASAX, os valores das séries temporais são clusterizados para a busca das regiões simbólicas com base na distribuição dos dados, o produto da clusterização são os centroides

(θ) que representam os centros das regiões simbólicas, cada região simbólica associasse a uma letra do alfabeto correspondente a discretização. Com base nos centros das regiões é possível calcular os pontos de quebra (β) com base na distância média entre dois centros vizinhos, conforme definido na equação 2.

$$\beta_i = \frac{\theta_{i-1} + \theta_i}{2} \tag{2}$$

3.5 Aproximação Agregada Simbólica por Zalewski et. al.

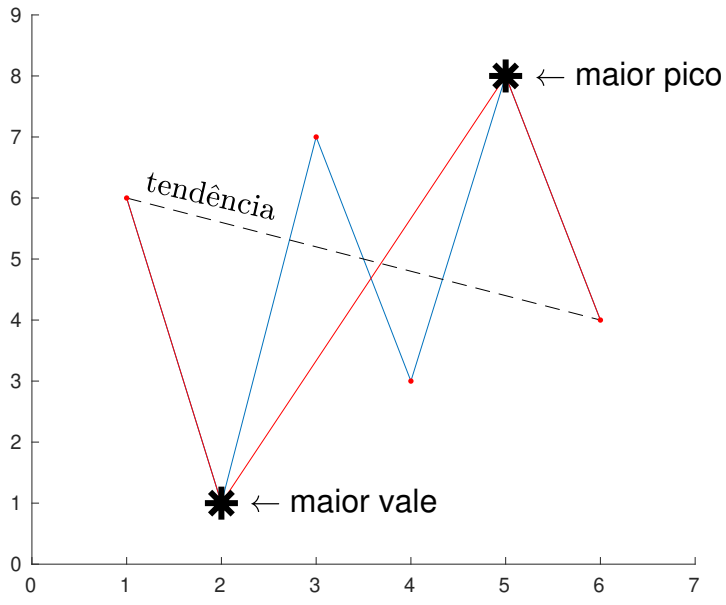
ZALEWSKI et al. propôs três estratégias de simbolização para representar uma série temporal (ZALEWSKI et al., 2012; ZALEWSKI et al., 2012; ZALEWSKI et al., 2013). Inicialmente, essa representação reduz a dimensionalidade por meio de uma estratégia semelhante ao PAA. No entanto, os valores resultantes dessa redução, não são simbolizados assumindo distribuição gaussiana da coordenada vertical, como no SAX. Ao contrário, os valores resultantes da redução da dimensionalidade passam pela diferenciação de primeira ordem, que consiste na diferença entre observações consecutivas em uma série temporal. Diferenciação de primeira ordem é muito utilizada para a remoção de tendências e ciclos em séries temporais (HYNDMAN; ATHANASOPOULOS, 2014).

Foram adotadas por ZALEWSKI et al. as seguintes formas de simbolização de valores contínuos, a saber, Discretização por Largura Igual (do inglês Equal Width Discretization - EWD) que divide os valores observados das variáveis em k intervalos, sendo k um parâmetro definido pelo usuário; Discretização de Frequência Igual (do inglês Equal Frequency Discretization - EFD) que divide os valores observados em k intervalos contendo m/k instâncias, sendo m a quantidade total de instâncias e k um parâmetro definido pelo usuário (DOUGHERTY et al., 1995; KOTSIANTIS; KANELLOPOULOS, 2006); Discretização por Valores Fixos e Iguais (do inglês Equal Fixed-Values Discretization - EFVD), a partir dos valores mínimo e máximo da série temporal são criadas k intervalos de mesmo tamanho, sendo k um parâmetro definido pelo usuário. É importante citar que a Discretização por Largura Igual pode ser suscetível a valores discrepantes (do inglês outliers) distorcendo os intervalos.

3.6 Aproximação Simbólica baseada em Tendência

Tendência é uma informação importante para determinados conjunto de séries temporais. Essa informação pode auxiliar no cálculo de similaridade entre séries temporais minimizando resultados indesejados. Aproximação Simbólica baseada em Tendência (do inglês Trend-based Symbolic approXimation - TSX) (LI; ZHANG; YANG, 2012) realiza a redução de dimensionalidade com base no PAA, porém além do valor médio de cada segmento extrai também a tendência relacionada ao maior pico e vale do segmento. A linha de tendência do segmento corresponde a reta que liga o primeiro valor do segmento ao último. O maior pico e vale do segmento tem relação com a tendência do segmento, neste caso é calculado a distância dos valores do segmento com a reta que liga o primeiro e último. Esse cálculo não é aplicado para o primeiro e último valor do segmento por questões obvias, como ilustrado na figura na figura 13. Por fim, o valor médio e as tendências são simbolizadas de acordo com pontos de parada equiprováveis para o valor médio e para os ângulos.

Figura 13 – Representação do TSX



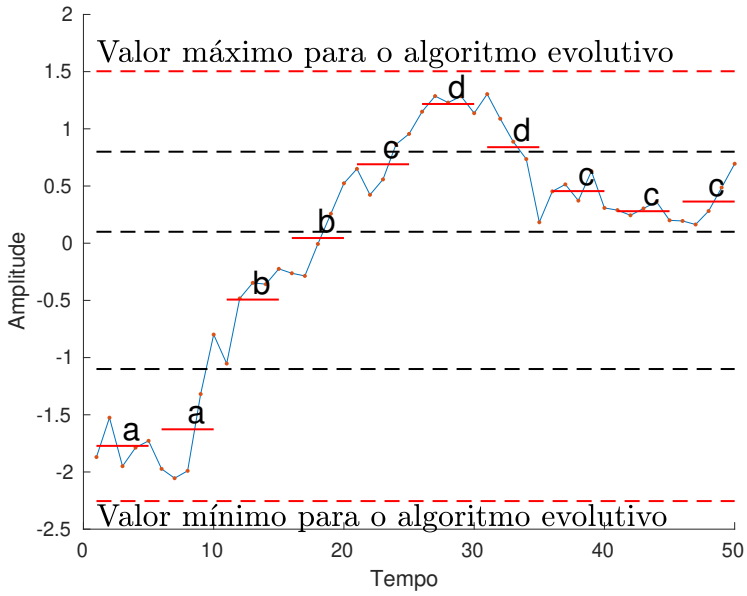
Fonte: Henrique Passos, 2018

3.7 Aproximação Agregada Simbólica baseado em Algoritmos Genéticos

A definição adequada das regiões a serem simbolizadas são fundamentais para o sucesso das representações de séries temporais simbólicas, no entanto, estas são altamente

relacionada com a distribuição dos dados a serem representados. Aproximação Agregada Simbólica baseado em Algoritmos Genéticos (do inglês Genetic Algorithms-Based Symbolic Aggregate approXimation - GASAX) (FUAD, 2012b) realiza a busca por pontos de quebra que melhor representam a série temporal independente da sua distribuição. No GASAX os possíveis pontos de quebras são codificados no cromossomo, os valores dos cromossomos variam dentro de um valor de mínimo é máximo. Esses pontos de quebra são utilizados no SAX para a simbolização dos dados, ao invés dos pontos de quebra que define regiões equiprováveis para a distribuição gaussiana. A figura 14 ilustra a representação do GASAX.

Figura 14 – Representação do GASAX



Fonte: Henrique Passos, 2018

3.8 Aproximação Agregada Simbólica baseado em Evolução Diferencial

Aproximação Agregada Simbólica baseado em Evolução Diferencial (do inglês Differential Evolution-Based Symbolic Aggregate approXimation - DESAX) (FUAD, 2012a) é uma proposta de representação de séries temporais baseada no GASAX que busca pontos de quebra para melhor simbolizar uma série temporal independente da sua distribuição. No caso do DESAX a busca é realizada por meio do algoritmo de evolução diferencial.

Similar ao GASAX, o DESAX busca os melhores pontos de quebra dentro de um intervalo definido pelo seu valor mínimo e máximo. O DESAX é proposto como uma evolução do GASAX e alcançou resultados relevantes comparado com o GASAX, ao

contrário do SAX que precisa que séries temporais tenham média zero e desvio padrão um, no DESAX essa normalização não é necessária. Fuad (2012a) apresentaram o conceito de não normalização da série temporal na representação. Essa não normalização mantem os valores originais coletados, os quais mantém o formato da série temporal e sua amplitude.

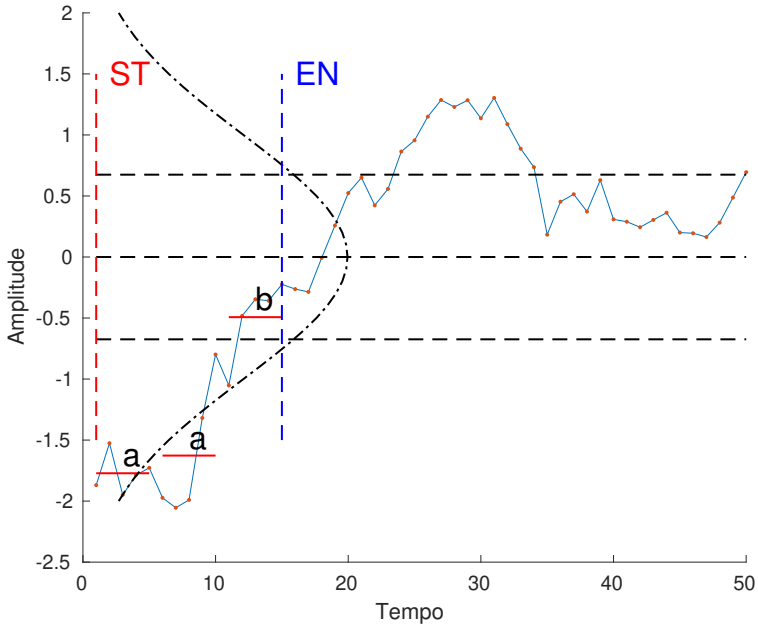
3.9 Aproximação Agregada Simbólica baseado em Variância

A maioria das representações simbólicas de séries temporais são baseadas no SAX e adotam segmentos de tamanho fixo em toda a série temporal, o qual pode não ser tão eficiente em termos de representação. O tamanho dos segmentos está diretamente ligado com a redução de dimensionalidade, isso porque cada segmento é aproximado por um único valor e quanto maior o tamanho do segmento maior a redução da quantidade de dados. Aproximação Agregada Simbólica baseado em Variância (do inglês Variance-Wise-Based Symbolic Aggregate approXimation - VWSAX) (SUN et al., 2012) define o tamanho dos segmentos de acordo com a variância dos dados. Segmentos de alta variância tendem a ter baixa redução de dimensionalidade dos dados e os segmentos de baixa variância tendem a ter alta redução de dimensionalidade dos dados. Os intervalos no VWSAX são definidos com base na análise iterativa da variância da série temporal. Todos os intervalos têm variância menor do que um limiar como ilustrado na figura 15, sendo ST e EN representando início e fim de cada segmento.

3.10 Aproximação Agregada Simbólica Melhorada

Diversas propostas de representações de séries temporais baseadas no SAX têm sido propostas na literatura, porém a representação realizada pelo SAX é destinada as séries temporais que possuem uma distribuição altamente gaussiana. Neste caso, a simbolização dos dados no SAX é baseada em regiões simbólicas equiprováveis. Representações como o ASAX, GASAX e GASAX, visam definir as regiões simbólicas que melhor representem os dados independente da sua distribuição, essas regiões são definidas com base em técnicas de agrupamento e busca por algoritmos evolutivos. Aproximação Agregada Simbólica Melhorada (do inglês Enhanced Symbolic Aggregate approXimation - ENSAX) (BARNAGHI; BAKAR; OTHMAN, 2012; BARNAGHI; BAKAR; OTHMAN, 2013) visa

Figura 15 – Representação do VWSAX



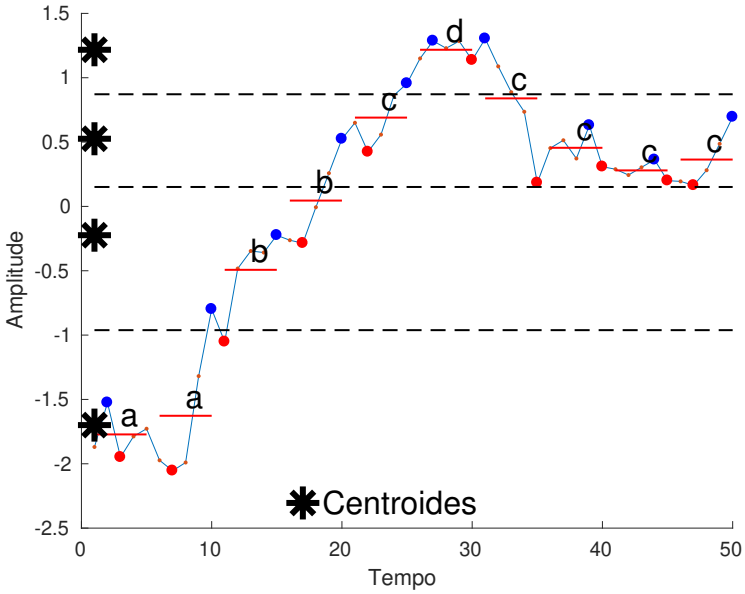
Fonte: Henrique Passos, 2018

melhorar simbolização dos dados ao levar em consideração os valores de mínimo, média e máximo dos segmentos. No entanto, esses valores não são utilizados diretamente na simbolização dos dados como no caso ESAX. No ENSAX cada segmento gera um vetor contendo os valores de mínimo, média e máximo. Os vetores gerados são agrupados por meio de uma técnica de agrupamento. Cada grupo gera um vetor protótipo. Ao invés de definir regiões simbólicas, o ENSAX simboliza os dados com base na similaridade com os vetores protótipo, como ilustrado na figura 16.

3.11 1D-Aproximação Agregada Simbólica

Durante a representação das séries temporais há transformações que enfatizam algumas características em detrimento de outras, essa é uma das maiores motivações para a criação de novas representações de séries temporais que buscam manter certas informações das séries temporais originais. No SAX a redução da dimensionalidade substitui os valores de cada segmento por uma aproximação feita pelo valor médio do segmento. Isso faz com que informações como a tendência sejam suprimidas. 1D-Aproximação Agregada Simbólica (do inglês 1D-Symbolic Aggregate approXimation - 1D-SAX) (SIMON et al., 2013) tem o objetivo de manter a informação de tendência dos segmentos, isso porque as representações de séries temporais baseadas no SAX podem não levar em consideração esta informação.

Figura 16 – Representação do ENSAX



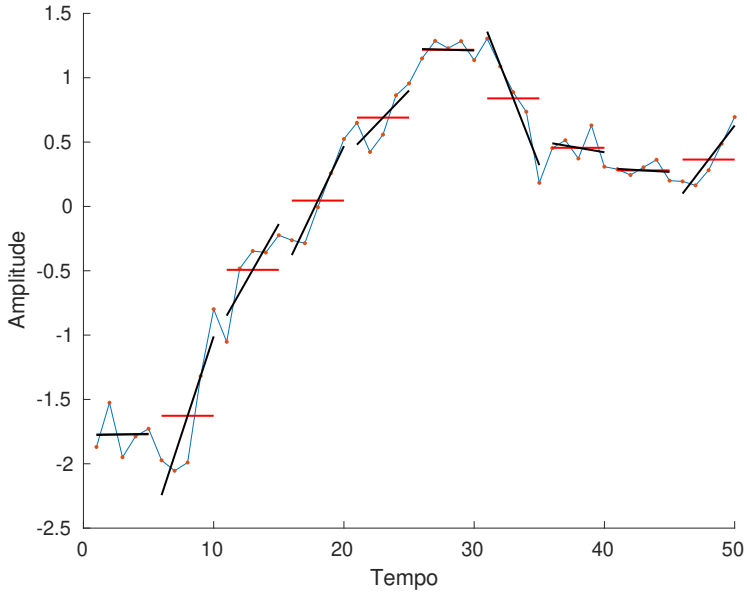
Fonte: Henrique Passos, 2018

No 1D-SAX a tendência é extraída com base no coeficiente linear da reta que melhor aproxima valores do segmento, como ilustrado na figura 17. Na versão original do 1D-SAX, esse coeficiente é quantizado e combinado com o valor médio do segmento gerando um único símbolo. Nesse trabalho não foi adotado a combinação do coeficiente linear com o valor médio transformando em um único valor, eles se mantiveram inalterados com um par de valores que representam os segmentos do 1DSAX. Vale lembrar que a tendência tem sua importância no cálculo de similaridade entre séries temporais, visto que segmentos representados pelo mesmo valor simbólico podem ter tendências distintas e nesse contexto o 1D-SAX tem a capacidade de distingui-los.

3.12 Aproximação Agregada Simbólica baseado em Mudança Aleatório

Várias representações de séries temporais simbólicas têm sido propostas com o objetivo de promover melhorias na sua capacidade de representação, visto que detalhes na redução de dimensionalidade ou até mesmo a forma de simbolizar os dados traz benefícios para a representação. Em relação a simbolização dos dados, o maior objetivo das representações é buscar os pontos de quebra que delimitam as regiões de atuação de cada símbolo que melhor iram simbolizar a série temporal com uma certa distribuição. Entretanto, esse tipo de divisão gera separações abruptas entre as regiões, levando em consideração que um valor perto de um ponto de quebra poderia ser simbolizado por uma

Figura 17 – Representação do 1D-SAX



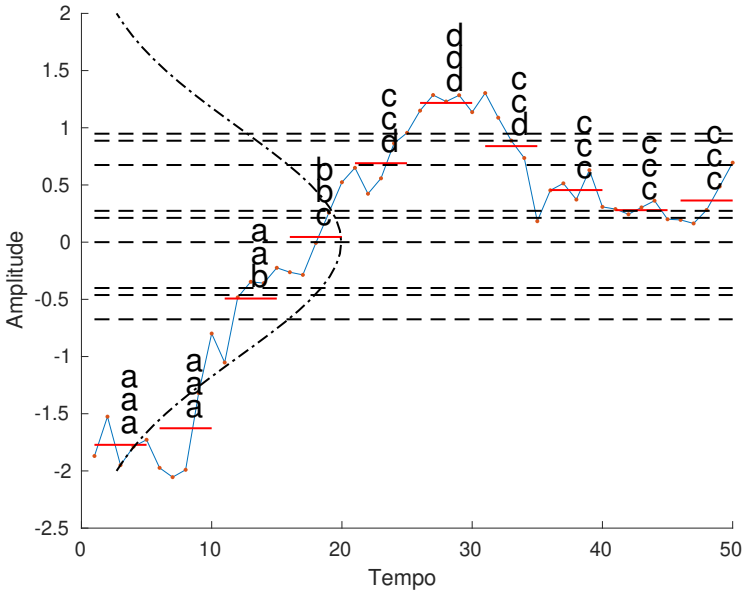
Fonte: Henrique Passos, 2018

outra letra do alfabeto. Aproximação Agregada Simbólica baseado em Mudança Aleatório (do inglês Random Shifting-Based Symbolic Aggregate approXimation - RSAX) (BAI et al., 2013) visa criar separações entre as regiões simbólicas que não sejam abruptas permitindo com que valores próximos dos pontos de quebra possam ser representados por mais de um símbolo. Neste caso, todos os pontos de quebra sofrem um deslocamento simultâneo possibilitando um segmento seja representado por mais de um símbolo. Os pontos de quebra podem ser deslocados tanto positivamente quando negativamente, a amplitude do deslocamento corresponde ao tamanho da menor região simbólica. Esse deslocamento pode ser realizado mais de uma vez gerando assim uma variação na simbolização dos valores próximos aos pontos de quebra, como ilustrado na figura 18.

3.13 Aproximação Agregada Simbólica baseado em Distância de Tendência

Em séries temporais a tendência pode ser uma informação valiosa como no cálculo de similaridade entre duas séries temporais. O 1D-SAX agrega a informação do coeficiente linear dos valores de cada segmento visando distinguir segmentos com mesmo valor médio por meio do coeficiente linear. Com base no princípio de manter a tendência dos segmentos a Aproximação Agregada Simbólica baseado em Distância de Tendência (do inglês Symbolic Aggregate approXimation-Trend Distance - SAXTD) (SUN et al., 2014) reduz a dimensionalidade e simboliza os dados semelhante ao SAX, porém na etapa de

Figura 18 – Representação do RSAX



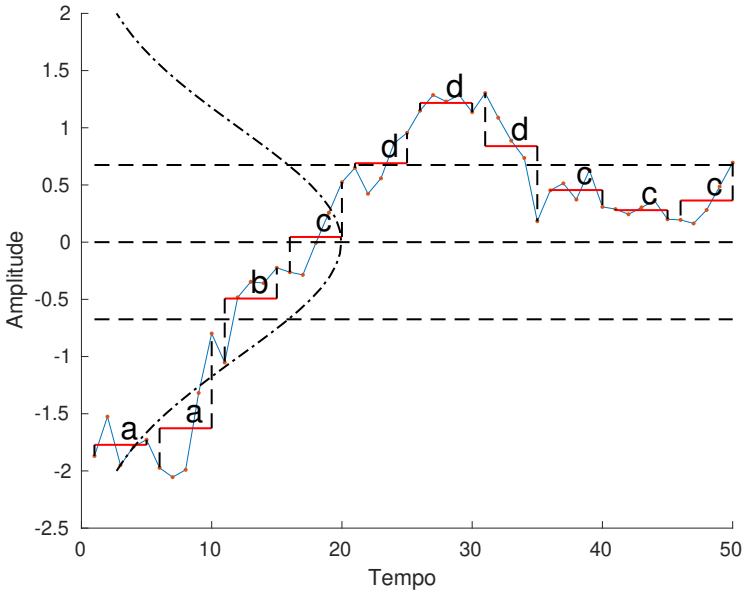
Fonte: Henrique Passos, 2018

redução de dimensionalidade, para cada segmento um valor que representa a sua tendência é extraído. O seu cálculo de tendência é consideravelmente intuitivo comparado com o 1D-SAX, no SAXTD a tendência de cada segmento é calculada com base na diferença entre o primeiro e último valor do segmento. Esse cálculo não leva em consideração a tendência global dos valores no segmento. Apesar do cálculo de tendência do SAXTD ser diferente do 1D-SAX o intuito de utilizar essa informação seque o mesmo princípio, isto é, possibilitar a distinção de segmentos com mesmo valor médio na análise de séries temporais. A figura 19 ilustra como o SAXTD calcula os padrões de tendência empregado na representação.

3.14 Aproximação Simbólica baseada em Tendência por Parte

Conforme já mencionado, a tendência é bastante importante para a análise de séries temporais como no cálculo de similaridade entre séries temporais. As representações de séries temporais extraem informações sobre a tendência das séries com o intuito de distinguir segmentos simbolizados pela mesma letra do alfabeto, mas com tendências distintas. Uma das formas de representar a tendência na literatura é através do coeficiente linear, porém tendência com ângulos próximos a $\pm 90°$ dificultam cálculos que levam em consideração as propriedades da tangente. Aproximação Simbólica baseada em Tendência por Parte (do inglês Symbolic Piecewise Trend Approximation - SPTA) (HATWAR; BADHIYE,

Figura 19 – Representação do SAXTD



Fonte: Henrique Passos, 2018

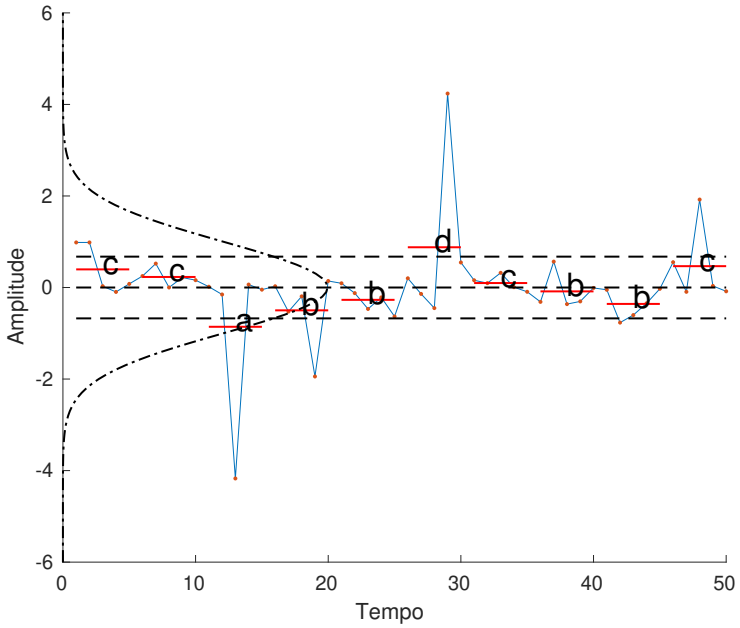
2015) expressa a tendência das séries temporais pela razão entre dois pontos consecutivos conforme equação 3, diante da série temporal $X = X_1, X_2, \ldots, X_N$ de tamanho N em que X_i é o i-ésimo valor da série temporal assumindo $\{i \in \mathbb{N} \mid 2 \leq i \leq N\}$, no SPTA essa razão entre dois pontos consecutivos resulta na série temporal normalizada sem tendência global $\hat{X} = \hat{X}_1, \hat{X}_2, \ldots, \hat{X}_{N-1}$. Essa razão é mais adequada para a representação da tendência dos dados, pelo fato da magnitude da razão representar o grau de variação da tendência e o seu sinal ($\pm$) representar sua direção (DAN et al., 2013). Um exemplo do SPTA é ilustrado na figura 20.

$$\hat{X}_i = \frac{X_i - X_{i-1}}{X_{i-1}} \tag{3}$$

3.15 Aproximação Constante por Parte Adaptativa

Para a redução de dimensionalidade muitas representações de séries temporais adotam o PAA, por ter alcançado resultados relevantes em diversas aplicações. Entretanto, a geração de segmentos de mesmo tamanho pode não ser a melhor solução para algumas séries temporais. Aproximação Constante por Parte Adaptativa (do inglês Adaptive Piecewise Constant Approximantion - APCA) (KEOGH et al., 2001b) busca reduzir a dimensionalidade das séries temporais por meio de segmentos de tamanho variado, cujo

Figura 20 – Representação do SPTA



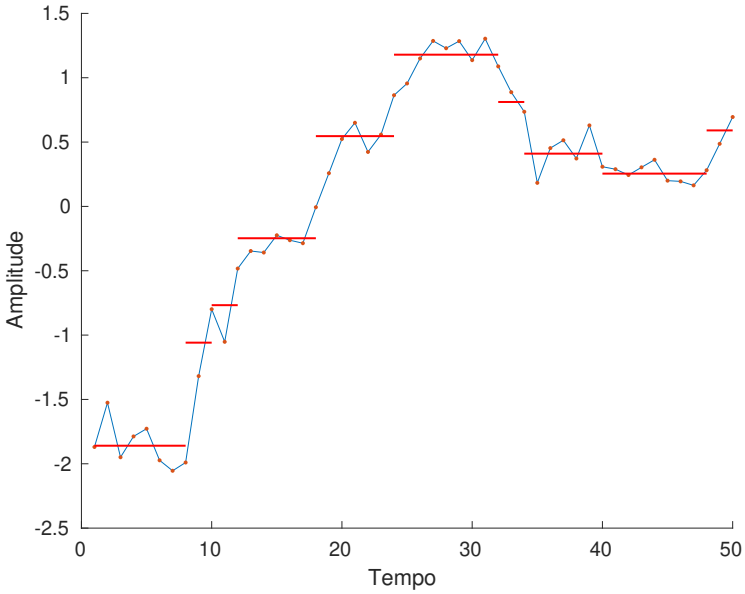
Fonte: Henrique Passos, 2018

valor médio melhor aproxima o segmento, como ilustrado na figura 21. A quantidade de segmentos é um parâmetro da representação e não mais o tamanho dos segmentos como no PAA. O APCA busca definir o tamanho dos segmentos minimizando o erro quadrático médio entre a aproximação e a série temporal original. Neste caso, a configuração de segmentos com o erro quadrático médio é solução da representação. Essa representação foi proposta por Keogh et al. (2001b) e obtiveram resultados relevantes na indexação de séries temporais.

3.16 Aproximação Agregada Simbólica Modificada

As novas propostas de representação de séries temporais simbólicas buscam encontrar segmentos de tamanho variado visando melhorar a redução de dimensionalidade, pontos de quebra que maximizam a representação dos dados independente da sua distribuição, informações que agregam valor na representação, como por exemplo, informação de tendência, dentre outras propostas. Aproximação Agregada Simbólica com K-Médias (do inglês Symbolic Aggregate approXimation K-Means - SAXKM) e Aproximação Agregada Simbólica Estendida com K-Médias (do inglês Extended Symbolic Aggregate approXimation K-Means - ESAXKM) (PASSOS et al., 2017) são representações de séries temporais que buscam encontrar os pontos de quebra visando melhor simbolizar os dados. Neste

Figura 21 – Representação do APCA



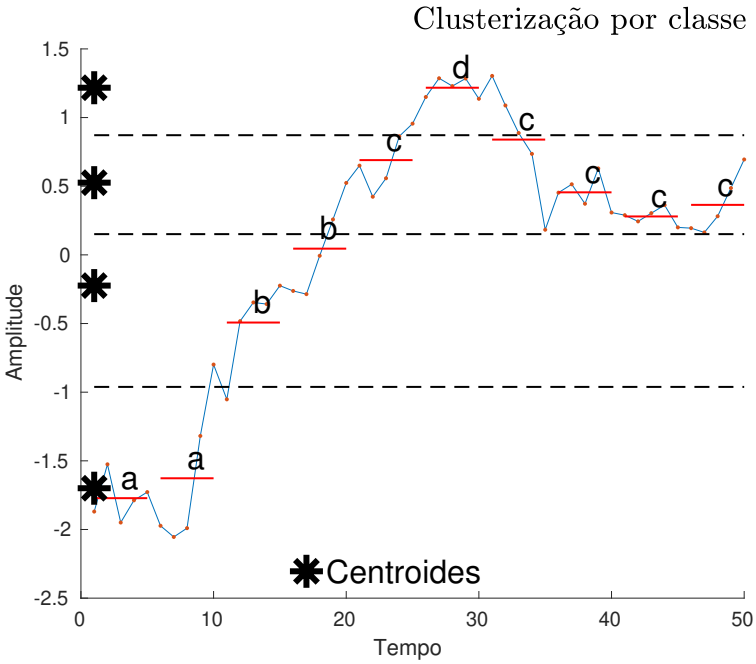
Fonte: Henrique Passos, 2018

caso, os pontos de quebra são provenientes de técnica de agrupamento dos dados antes da redução de dimensionalidade, com isso há mais informação que pode ajudar a definir o cálculo dos pontos de quebra. Outra particularidade dessas representações reside na técnica de agrupamento que é aplicada classe a classe, pois se alguma classe tiver uma distribuição diferente da distribuição do conjunto de dados como um todo, a sua representação pode ser afetada. Diversas representações apresentadas anteriormente mostraram que os pontos de quebra têm grande influência na representação dos dados. Tanto no SAXKM quanto no ESAXKM os pontos de quebra resultantes do agrupamento são utilizados para a simbolização dos dados. O SAXKM representa cada segmento pelo seu valor médio e o ESAXKM representa cada segmento pelos seus valores médio, mínimo e máximo. Uma versão diferente do SAXKM é a Representação k-Médias (do inglês Representation K-Means - R-Kmeans) que é focado na simbolização dos dados. No R-Kmeans a série temporal é normalizada com média zero e desvio padrão um, os valores da série temporal são agrupados visando encontrar os pontos de quebra como no ASAX e posteriormente os dados são simbolizados.

3.17 Comparação das técnicas de representações de séries temporais

O entendimento sobre as particularidades entre as representações é importante para analisar o real benefício da aplicação de uma representação em uma série temporal.

Figura 22 – Representação do SAXKM



Fonte: Henrique Passos, 2018

Por exemplo, o ESAX é uma representação que apresentou resultados relevantes quando a série temporal analisada consiste de valores de ações da bolsa de valores (LKHAGVA; SUZUKI; KAWAGOE, 2006b). Neste caso, os valores de máximo e mínimo são importantes para caracterizar uma série temporal. Outras representações de séries temporais que tem objetivos semelhantes também podem alcançar desempenho similar. As representações de séries temporais têm suas especificidades e é isso que esse capitulo tenta sumarizar. A tabela 1 apresenta de forma sucinta os principais aspectos entre as representações de séries temporais descritas acima. Essa tabela apresenta a normalização, a segmentação, a redução de dimensionalidade e as formas de simbolização empregadas pelas representações. A simbolização e redução de dimensionalidade foram as etapas dentro das representações de séries temporais com maior quantidade de variações propostas na literatura. Essas duas etapas dentro de uma representação desempenham papel primordial na transformação dos dados, a ineficiência em termos da representação na transformação dos dados implica em baixo desempenho dessa representação.

Na tabela 1 existem diversas nomenclaturas que descrevem as diferentes formas de representações de séries temporais, a seguir os seus significados: NMDP) significa que a série temporal é normalizada com média zero e desvio padrão um; STU) segmentação de tamanho uniforme; STVE) segmentação de tamanho variado, o qual é definido minimizando

Tabela 1 – Comparação das representações de séries temporais

Representação	Normalização	Segmentação	Redução	Simbolização
1DSAX	NMDP	STU	RDMCL	SBGS
APCA	NMDP	STVE	RDM	-
ASAX	NMDP	STU	RDM	SBKM
DESAX	-	STU	RDM	SBGA
EFD	NMDP	STU	RDMDPO	SBCF
EFVD	NMDP	STU	RDMDPO	SBCV
ENSAX	NMDP	STU	RDM	SBGS
ESAX	NMDP	STU	RDMMM	SBGS
ESAX-Kmeans	NMDP	STU	RDMMM	SBKM
EWD	NMDP	STU	RDMDPO	SBCL
GASAX	-	STU	RDM	SBGA
R-Kmeans	NMDP	STU	RDM	SBKM
PAA	NMDP	STU	RDM	-
RSAX	NMDP	STU	RDM	SBGSMA
SAX	NMDP	STU	RDM	SBGS
SAX-Kmeans	NMDP	STU	RDM	SBKM
SAXTD	NMDP	STU	RDMTD	SBGS
SPTA	NMDP	STU	RDMRT	SBGS
VWSAX	NMDP	STVV	RDM	SBGS

Fonte: Henrique Passos, 2018

o erro quadrático médio; STVV) segmentação de tamanho variado, o qual é definido com base na variância; RDM) redução de dimensionalidade com aproximação pelo valor médio; RDMCL) redução da dimensionalidade com aproximação pelo valor médio e mantem o valor do coeficiente linear do segmento; RDMDPO) redução de dimensionalidade com aproximação pelo valor médio e posteriormente há a execução de diferenciação de primeira ordem; RDMMM) redução de dimensionalidade com aproximação pelo valor médio e mantem os valores mínimo e máximo do segmento; RDMTD) redução de dimensionalidade com aproximação pelo valor médio e mantem os valores sobre a tendência do segmento (valor inicial menos final de cada segmento); RDMRT) redução de dimensionalidade com aproximação pelo valor médio e mantem os valores sobre a razão de tendência do segmento; SBGS) simbolização é baseada na divisão da gaussiana em área equiprováveis (semelhantes ao SAX); SBKM) simbolização é baseada nos centroides provenientes da técnica de agrupamento dos dados; SBGA) simbolização é baseada na divisão das regiões delimitadas por algoritmos evolutivos; SBCF) simbolização é baseada na compartimentalização dos dados por frequência dos valores; SBCV) simbolização é baseada na compartimentalização dos dados por regiões iguais; SBCL) simbolização é baseada na compartimentalização dos dados por largura dos valores; SBGSMA) simbolização é baseada na divisão da gaussiana em área equiprováveis em múltiplas amostras;

4 Algoritmos Evolutivos

Na natureza evoluir está relacionado a capacidade de sobreviver e prosperar. O mecanismo de evolução inspirou diversos algoritmos evolutivos, criando uma subárea chamada computação evolucionaria. Na natureza através dos seus mecanismos evolutivos uma série de problemas complexos são solucionados geração após geração (iterativamente), esse é um mecanismo de otimização espontâneo que a natureza emprega para resolver problemas. Do ponto de vista computacional, dado um problema de otimização e uma função de *fitness* que medi a qualidade da solução, o algoritmo evolutivo de forma geral busca por soluções aproximadas em problemas de otimização (LANGDON; POLI, 2013). Dado que as representações simbólicas de séries temporais possuem diversos parâmetros a serem definidos e que cada representação possui vantagens e desvantagens, esta dissertação empregou algoritmos evolutivos para selecionar os parâmetros das representações simbólicas e qual(is) representações são mais adequadas um ensemble visando atacar um problema específico. Em função este capítulo apresenta uma descrição detalhada dos algoritmos evolutivos utilizados nesta dissertação.

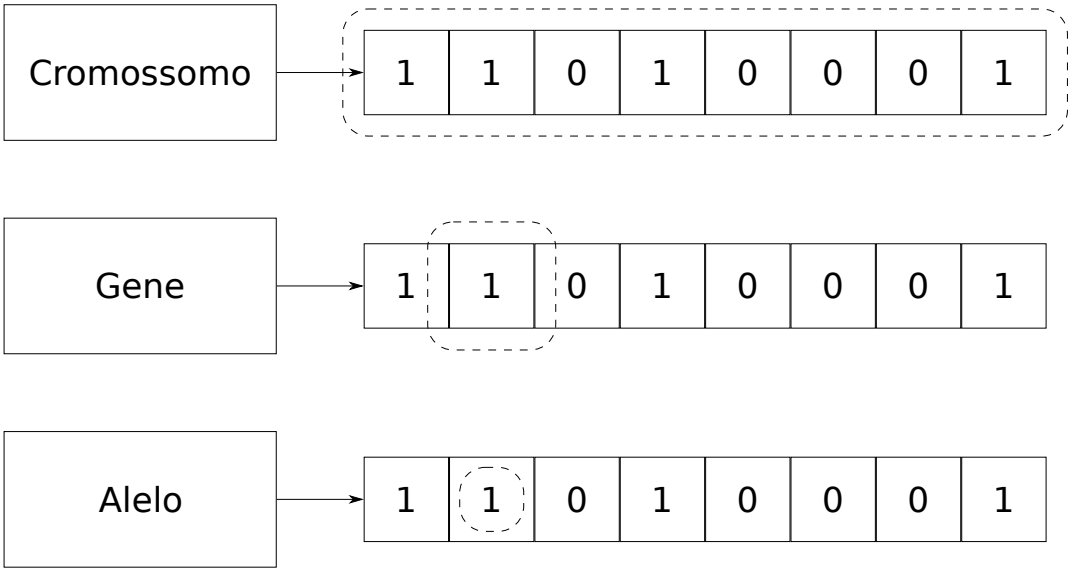
4.1 Algoritmo Genético

Algoritmo genético (do inglês Genetic Algorithm - GA) foi proposto por HOLLAND na década de 1960 e desenvolvido pelos seus estudantes e colegas da Universidade de Michigan nas décadas de 1960 e 1970. Em 1975 no livro "*Adaptation in Natural and Artificial Systems*" o algoritmo genético foi descrito como uma abstração da evolução biológica e um arcabouço teórico foi apresentado (MITCHELL, 1998). GAs tem como objetivo de buscar por soluções aproximadas em problemas de otimização.

O GA apresentado por HOLLAND consiste em evoluir uma população recorrendo a mecanismo de seleção natural juntamente com operadores inspirados na genética, tal como recombinação (do inglês crossover) e mutação. A população corresponde ao conjunto de soluções candidatas, onde cada solução candidata ou instância é representada por cromossomo, sendo que cada posição do cromossomo é chamada de gene e cada valor é chamado de alelo, como ilustrado na figura 23. O operador de seleção tem um papel importante no algoritmo. Há vários mecanismos de seleção, tais como, roleta e torneio, que consistem

de mecanismos de seleção não determinístico. Nestes os indivíduos de maior desempenho com base na função objetivo têm a maior probabilidade de ser selecionado para compor a próxima geração. A recombinação é a troca de alelos entre cromossomos, que corresponde a uma imitação do processo de recombinação biológica entre dois cromossomos simples. A recombinação permite a geração de novos indivíduos que herdam as características dos pais, com isto é possível explorar novas regiões do espaço de busca pela combinação dos valores presentes na população. As mutações geram variações no conjunto de genes da população, isso introduz e contribui para a diversidade genética da população. Essas variações são de natureza aleatória proporcionando a exploração de novas regiões no espaço de busca (MITCHELL, 1998).

Figura 23 – Nomenclaturas algoritmo genético



Fonte: Henrique Passos, 2018

No algoritmo genético, os genes podem ser codificados com valores contínuos e/ou discretos e para cada gene existe um intervalo de valores possíveis de serem assumidos pelo mesmo. Após a definição da codificação a ser utilizadas nos genes, é realizada a inicialização dos cromossomos que irão compor a população inicial. A população será avaliada por meio da função objetivo (do inglês *fitness*), que mede o grau de adaptabilidade do cromossomo. O algoritmo 1 descreve os principais passos do algoritmo genético. Os parâmetros de entrada para o algoritmo 1 são: o tamanho da população que representa o conjunto de soluções candidatas passíveis de ser evoluídas pelo algoritmo genético e a função objetivo que avalia o cromossomo de acordo com o problema a ser otimizado.

Algoritmo 1 Descrição do GA

```
1: para todo cromossomo faça
2:     inicializar cromossomo
3: fim para
4: enquanto não atingir o máximo de épocas ou o critério de erro mínimo faça
5:     calcular o valor do desempenho dos cromossomos
6:     selecionar os cromossomos para próxima geração
7:     recombinar os cromossomos
8:     realizar mutações nos cromossomos
9: fim enquanto
```

Fonte: Henrique Passos, 2018

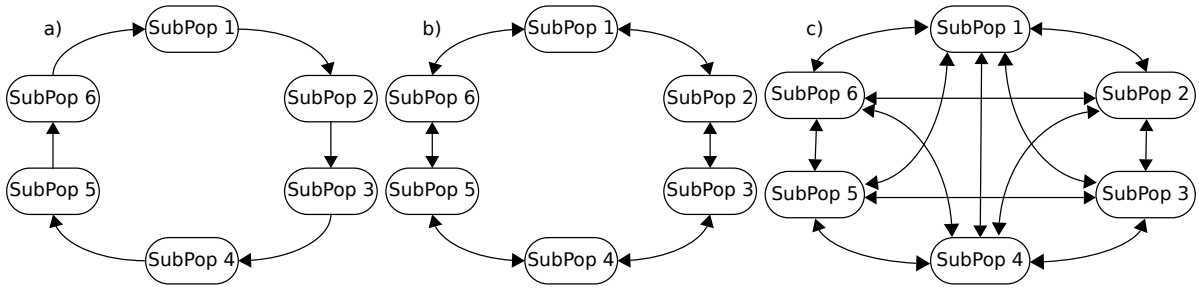
4.1.1 Algoritmo genético com população múltipla

A biologia evolucionista mostra que grandes populações centrais estáveis exercem forte influência na homogeneização, visto que novas mutações são diluídas na grande massa populacional. Em função disso as diferenças entre os indivíduos da população passam a ser rara, porém em pequenas populações a pressão seletiva se intensifica e variações favoráveis são herdadas rapidamente (GOULD, 1977). Analogamente, nos algoritmos genéticos os operadores genéticos são empregados tradicionalmente sobre uma única população. A evolução uni-populacional pode limitar a exploração do espaço de busca. A variação do algoritmo genético com múltiplas populações (do inglês multiple populations genetic algorithm - MPGA) (STARKWEATHER; WHITLEY; MATHIAS, 1990; MÜHLENBEIN; SCHOMISCH; BORN, 1991) tem o objetivo de minimizar essa limitação, neste caso exploração do espaço de busca pode se concentrar em locais distintos no espaço de busca (TOLEDO et al., 2005).

Outro ponto a ser considerado no algoritmo genético com população múltipla é a possibilidade da migração de bons indivíduos entre as subpopulações visando a melhoria na evolução dos descentes. No entanto, não é regra que a migração seja realizada somente com bons indivíduos. A regra de origem e destino na migração dos indivíduos entre populações precisa ser definida, existem diversas topologias migratórias como: anel, vizinhança e irrestrita. Na topologia anel as migrações são unidirecionais entre as subpopulações adjacentes e a última subpopulação se interliga a primeiro formando uma topologia em forma de anel. Semelhante a topologia anel, a topologia de vizinhança tem o fluxo migratório em anel, porém essa migração é bidirecional. A topologia que tem a maior

liberdade na migração é a irrestrita, nessa topologia a migração ocorre de todas para todas as subpopulações. A figura 24 ilustra os três exemplos de topologias.

Figura 24 – Topologia da migração: a) anel, b) vizinho e c) irrestrita



Fonte: Henrique Passos, 2018

4.2 Algoritmo de Evolução Diferencial

Algoritmo de Evolução Diferencial (do inglês Differential Evolution - DE) (STORN, 1996; STORN; PRICE, 1997) possibilita resolver problemas de otimização baseados em valores contínuos. Na evolução diferencial as possíveis soluções para o problema da função objetivos são vetores de valores contínuos chamados de agentes, os agentes $x_{i,G},\ i = 1, 2, \ldots, NP$, sendo que NP está relacionado com o tamanho da população de agentes, esses agentes serão evoluídos ao decorrer de G gerações. A mutação é uma das operações exercida pela evolução diferencial que tem o objetivo de explorar novas regiões no espaço de busca. Na mutação três agentes são selecionados $r_1, r_2, r_3 \in \{1, 2, \ldots, NP\}$, sendo r_1, r_2, r_3 os índices dos agentes selecionados e o novo agente é gerado através da combinação dos três agentes selecionados, como na equação 4. O fator de mutação (F) na equação pondera a diferença dos agentes $(x_{r_{2,G}} - x_{r_{3,G}})$ e adicionada ao agente $x_{r_1,G}$. Esse fator normalmente assume valores entre $(0, 2)$, sendo zero a ausência de mutação e valores próximos a dois a ênfase na mutação do agente $x_{r_1,G}$. O resultado da mutação é o agente $v_{i,G+1}$.

$$v_{i,G+1} = x_{r_1,G} + F \times (x_{r_{2,G}} - x_{r_{3,G}}) \tag{4}$$

Após a mutação existe a recombinação dos agentes $v_{i,G+1}$ e $x_{ji,G}$. O parâmetro taxa de recombinação (do inglês Crossover Rate - CR) representa a probabilidade da recombinação do agente $v_{i,G+1}$ em $x_{ji,G}$, esse parâmetro assume valores entre (0, 1) representando a probabilidade de recombinação, a condição $U(0, 1) \leq CR$ representa a probabilidade

de recombinação. A condição $j = U(1, D)$ faz com que ao menos um valor do $v_{i,G+1}$ com D-dimensões seja recombinado. O agente $u_{i,G+1}$ corresponde a resultado da recombinação do agente $v_{i,G+1}$ com o agente $x_{ji,G}$, como definido na equação 5.

$$u_{ji,G+1} = \begin{cases} v_{ji,G+1} & \text{if } U(0,1) \leq CR \text{ or } j = U(1,D) \\ x_{ji,G} & \text{caso contrário} \end{cases}, \quad j = 1, 2, \ldots, D \tag{5}$$

Ao final, os agentes $u_{i,G+1}$ e $x_{ji,G}$ são avaliados pela mecanismo de seleção para a próxima geração, caso o agente gerado pela recombinação $u_{i,G+1}$ tenha desempenho melhor do que o agente $x_{ji,G}$ com base na função objetivo, esse agente gerado passa para a próxima geração. A convergência comumente é baseada na quantidade máxima de iterações ou erro mínimo alcançado pelo melhor agente. O algoritmo 3 apresenta de forma simplificada o DE.

Algoritmo 2 Descrição do DE

para todo agente **faça**
    inicializar agente
fim para
faça
    para todo agente **faça**
        $v_{i,G+1} = x_{r_1,G} + F \times (x_{r_2,G} - x_{r_3,G})$ ▷ mutação
        $u_{ji,G+1} = \begin{cases} v_{ji,G+1} & \text{if } U(0,1) \leq CR \text{ or } j = U(1,D) \\ x_{ji,G} & \text{caso contrário} \end{cases}$, $j = 1, 2, \ldots, D$ ▷ recombinação
    fim para
    se o desempenho de $u_{i,G+1}$ for maior do que $x_{ji,G}$ **então**
        substituir $u_{i,G+1}$ por $x_{ji,G}$
    fim se
enquanto não atingir o máximo de iterações ou o critério de erro mínimo

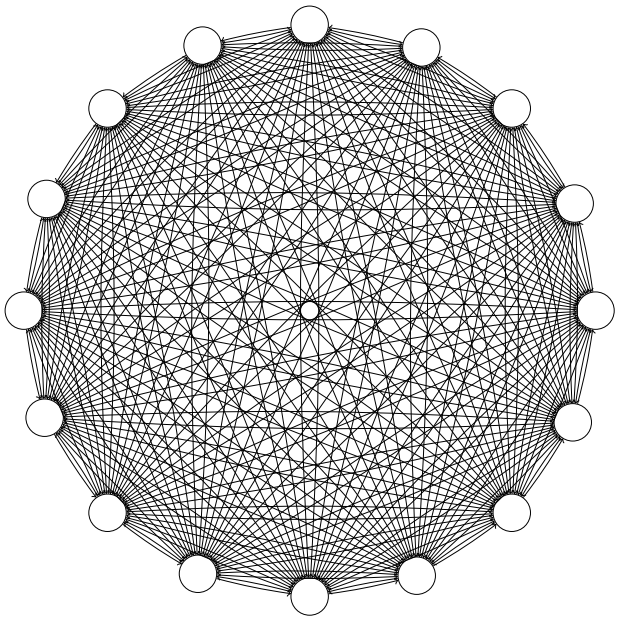
Fonte: Henrique Passos, 2018

4.3 Otimização por Enxame de Partículas

Otimização por Enxame de Partículas (do inglês Particle Swarm Optimization - PSO) (EBERHART; KENNEDY, 1995; KENNEDY; EBERHART, 1997; CLERC; KENNEDY, 2002) se baseia no comportamento de populações, esse comportamento coletivo foi aplicado a computação evolucionária. As partículas estão dispostas no espaço de busca

multidimensional, as suas posições representam as possíveis soluções para o problema definido na função objetivo. A variável responsável pela posição da partícula é comumente chamada de x, o i é o seu índice e o t o seu instante de tempo. Essas partículas se deslocam devido a influência de três forças: a inércia, o melhor resultado individual e global (vizinhança) das partículas. A inércia é a força que conserva a partícula a manter uma direção. A partícula tende a se movimentar para regiões que foram favoráveis ao problema, o melhor resultado já alcançado pela partícula tem influência no seu deslocamento futuro, essa variável é comumente chamada de *pbest*. De forma similar, o melhor resultado dentre todas as partículas influência as demais de forma geral, essa variável é comumente chamada de *gbest*. Essas variáveis são experiências individuais e coletivas das partículas. O *gbest* representa a influência da melhor partícula em uma vizinhança, essa vizinhança é definida através de uma topologia com: anel, altamente conectada, estrela, malha, arvore e etc. As diferentes topologias podem afetam a convergência (KENNEDY, 1999; MEDINA; PULIDO; RAMÍREZ-TORRES, 2009) e para esse trabalho a topologia adotada foi a altamente conectada, sendo que todos as partículas são vizinhas de todas as outras. A figura 25 é um exemplo de topologia de vizinhança altamente conectada, os círculos representam as partículas e as retas a conectividade topológica da vizinhança entre as partículas, por ser um grafo completo todas as partículas são reciprocamente vizinhas.

Figura 25 – Topologia de vizinhança altamente conectada



Fonte: Henrique Passos, 2018

Através de operações matemáticas, essas forças são combinadas e ponderadas resultando na variação da velocidade e direção das partículas, essa variação tem como objetivo explorar o espaço de busca por novas soluções. As variáveis estocásticas U_1 e U_2 são fruto da distribuição uniforme (0,1), tem o objetivo de adicionar uma variação na função de atualização de velocidade considerando a posição das melhores partículas *pbest* e *gbest*, essas variações simulam alguns comportamentos aleatórios presentes na população de partículas. φ_1 e φ_2 são fatores de aprendizado comumente definidos como 2, tanto $\varphi_1 U_1(0,1)$ quando $\varphi_2 U_2(0,1)$ assumem na média o valor igual a 1. Durante a execução do PSO uma partícula pode assumir valores infinitos tanto positivamente quando negativamente, com isso é necessário a definição valores de mínimo e máximo para cada dimensão da partícula. O algoritmo 3 apresenta de forma simplificada o PSO.

Algoritmo 3 Descrição do PSO

```
para toda partícula faça
    inicializar partícula
fim para
faça
    para toda partícula faça
        calcular o desempenho
        se o desempenho é melhor do que o pbest então
            o desempenho atual é o novo pbest
        fim se
    fim para
    escolher o melhor desempenho dentre todas partículas para ser a gbest
    para toda partícula faça
        calcular a nova velocidade
        atualizar a posição
    fim para
enquanto não atingir o máximo de iterações ou o critério de erro mínimo
```

Fonte: Henrique Passos, 2018

Após a definição dos melhores valores locais *pbest* e global *gbest*, a equação 6 é responsável por atualizar a velocidade e posição das partículas respectivamente, sendo i uma determinada partícula e t um instante de tempo. Essas equações definem a dinâmica de movimentação das partículas no espaço de busca multidimensional no PSO tradicional.

$$\begin{aligned} v_i^{t+1} &= v_i^t + \varphi_1 U_1(0,1) * (pbest_i - x_i^t) + \varphi_2 U_2(0,1) * (gbest - x_i^t) \\ x_i^{t+1} &= x_i^t + v_i^{t+1} \end{aligned} \tag{6}$$

Um dos pontos que ficam em aberto na formulação tradicional do PSO é a adequação da inercia das partículas no deslocamento do espaço de busca, a adequação da inercia está

diretamente ligado com o comportamento das partículas na exploração do espaço de busca. Shi e Eberhart (1998) propuseram a introdução de um novo parâmetro na formulação tradicional do PSO, o peso da inercia (w) que pondera a velocidade das partículas como apresentado na equação 7, o valor de w maiores do que 1 proporcionam uma maior exploração inicial do espaço de busca, o valor de w menos do que 1 contribui na exploração das regiões *pbest* e *gbest*.

$$v_i^{t+1} = wv_i^t + \varphi_1 U_1(0,1) * (pbest_i - x_i^t) + \varphi_2 U_2(0,1) * (gbest - x_i^t) \tag{7}$$

Outra variação do PSO semelhante ao peso da inercia é a inserção do fator de enxugamento (χ), esse fator altera o comportamento das partículas na convergência ao entorno dos melhores valores encontrados (KENNEDY; EBERHART, 1997; CLERC; KENNEDY, 2002). Com valores de $\phi < 4$ as partículas exploram a melhor solução alcançada no espaço de busca lentamente como se fosse um "espiral" não garantindo da convergência, esse tipo de parametrização foca na busca de soluções ao redor do melhor valor. Já assumindo valores $\phi > 4$ é praticamente o oposto, a convergência é direta, precisa e garantida. A equação 8 apresenta a atualização de velocidade com o fator de enxugamento. Por questões de conveniência, é comum no PSO os valores de φ_1 e φ_2 assumirem o mesmo valor, isso faz com que a atualização de velocidade das partículas tanto de *pbest* e *gbest* tenham o mesmo peso, na formulação do PSO com fator de enxugamento (do inglês Constricted Particle Swarm Optimization - CPSO) os parâmetros φ_1 e φ_2 são comumente iguais a 2.05, ϕ é igual a 4.1 e o fator de enxugamento χ assume o valor de 0.72984.

$$\begin{aligned} v_i^{t+1} &= \chi[v_i^t + \varphi_1 U_1(0,1) * (pbest_i - x_i^t) + \varphi_2 U_2(0,1) * (gbest - x_i^t)] \\ \chi &= \frac{2}{\left|2 - \phi - \sqrt{\phi^2 - 4\phi}\right|}, \\ \phi &= \varphi_1 + \varphi_2 \end{aligned} \tag{8}$$

O PSO também pode ser encontrado na versão binaria (do inglês Binary Particle Swarm Optimization - BPSO) (KENNEDY; EBERHART, 1997; CLERC; KENNEDY, 2002), os valores de x_i são bits que equivalem as variáveis da função objetivo no formato binário. O deslocamento das partículas no espaço de busca das variáveis binarias são baseados na inércia, a variável v_i que no PSO tradicional representa a velocidade das partículas, no PSO com codificação binaria é o termo de probabilidade de mudança do bit

para o valor um, um valor de probabilidade de mudança de um bit qualquer igual a 0.20 significa de esse bit tem 20% de chances de ser o valor um, a função sigmoide na equação 9 limita os valores de x_i^{t+1} entre zero e um. As alterações nos bits são responsáveis pela exploração do espaço de busca binário baseado na probabilidade.

$$\begin{aligned} x_i^{t+1} &= \begin{cases} 0, & \text{if } U(0,1) \geq S(v_i^{t+1}) \\ 1, & \text{if } U(0,1) < S(v_i^{t+1}) \end{cases}, \\ S(x) &= \frac{1}{1+e^{-x}} \end{aligned} \tag{9}$$

5 Ensemble de técnicas de representação simbólica

Em vez de se fixar no estudo de preditores operando isoladamente, a subárea de pesquisa conhecida como comitês de máquinas se ocupa do estudo de diferentes maneiras de se compor vários desses estimadores – denominados especialistas ou componentes e possivelmente oriundos de distintos métodos de aprendizado (não-)supervisionado – em um só arranjo conceitual. Uma primeira motivação em se trabalhar com comitês de máquinas é a de justamente eliminar a necessidade de se ter que configurar acertadamente parâmetros de controle associados aos indutores ou se estipular previamente qual deva ser a dimensão do modelo (máquina) resultante. Outra forte justificativa é que os módulos que compõem tais sistemas têm propensão a apresentar menor complexidade, facilitando o seu entendimento por parte do projetista, bem como uma possível modificação ou extensão a posteriori. Neste capítulo, é apresentado um breve panorama de comitê de máquina, mostrando as principais razões pelas quais um comitê é capaz de apresentar um desempenho superior a qualquer componente. Em seguida, é apresentado uma descrição detalhada da abordagem proposta nesta dissertação.

Uma série temporal estocástica consiste de uma sequência ordenada de observações medidas ao longo do tempo: $z = \{z_1, z_2, \cdots, z_N\}$, a qual corresponde a uma das possíveis trajetórias passíveis de serem geradas por um processo físico em observação, este denominado processo estocástico, o qual, por sua vez, pode ser encarado como um conjunto de variáveis aleatórias, uma para cada índice do tempo t inteiro. Cada valor observado de uma trajetória alude a um dos possíveis valores que poderiam ter sido observados, de acordo com a densidade de probabilidades da respectiva variável aleatória. As séries temporais podem ser originadas a partir de processos de natureza bastante diversa, razão pela qual elas recebem adjetivações específicas, como séries temporais financeiras, séries temporais biológicas e séries temporais climáticas.

Analise de séries temporais tem sido usada com sucesso em várias aplicações para identificar padrões temporais nos dados. O sinal de ECG ou EEG ou EMG pode ser considerado como uma série temporal. Uma questão chave no processamento eficaz e eficiente de séries temporais está ligada à sua representação sequencial. A representação de séries temporais pode ser realizada usando métodos no domínio tempo, frequência

e tempo-frequência. A transformada de Fourier realiza a representação no domínio da frequência, enquanto as Representações Simbólica no domínio do tempo.

A representação simbólica pode melhorar a análise de processos que são complexos e possivelmente caótico. Além disso, representação simbólica pode também reduzir a sensibilidade a ruído e melhorar bastante a eficiência computacional. No entanto, aspectos estruturais e paramétricos de projeto podem conduzir a uma degradação de desempenho. Tais aspectos dizem respeito principalmente à escolha a priori dos parâmetros das técnicas de representação simbólica, tais como, tamanho do alfabeto, tamanho do segmento, etc. Além disso, a escolha inadequada da técnica de representação simbólica para um problema especifico pode conduzir a um desempenho aguem daquele indicado na literatura. Assim, na ausência de uma metodologia sistemática e de baixo custo computacional, ensemble de técnicas de representação simbólica se apresentam como alternativas capazes de amenizar significativamente os efeitos negativos de especificações equivocadas de projeto. Mais ainda, a existência de configurações estruturais e paramétricas alternativas para as técnicas de representação simbólica acaba por sustentar a implementação de ensembles. Sendo assim, o objetivo deste capítulo é detalhar a abordagem proposta, que consiste em projetar um ensemble de técnicas de representação simbólica via algoritmo evolutivo.

5.1 Comitê de Máquinas

Em vez de se fixar no estudo de preditores operando isoladamente, a subárea de pesquisa conhecida como comitês de máquinas (KUNCHEVA, 2004; TRESP, 2001) se ocupa do estudo de diferentes maneiras de se compor vários desses estimadores – denominados especialistas ou componentes e possivelmente oriundos de distintos métodos de aprendizado (não-)supervisionado – em um só arranjo conceitual. Uma primeira motivação em se trabalhar com comitês de máquinas é a de justamente eliminar a necessidade de se ter que configurar acertadamente parâmetros de controle associados aos indutores ou se estipular previamente qual deva ser a dimensão do modelo (máquina) resultante. Outra forte justificativa é que os módulos que compõem tais sistemas têm propensão a apresentar menor complexidade, facilitando o seu entendimento por parte do projetista, bem como uma possível modificação ou extensão a posteriori.

No contexto supervisionado, duas classes de abordagens complementares de comitês de máquinas podem ser discriminadas na literatura (SHARKEY, 2012). A primeira classe se baseia mais explicitamente no preceito de "dividir para conquistar". Desse modo, uma tarefa complexa passa a ser decomposta em certo número de subtarefas, cada uma delas alocada a um subgrupo de especialistas. Uma forma de se modularizar a arquitetura do sistema é a de se definir, dinamicamente, regiões do espaço de entradas a serem atribuídas aos especialistas; assim, a decisão do grupo para uma dada amostra passa a ser incumbência somente daquele(s) especialista(s) alocado(s) à região a que ela pertence. Essa estratégia de particionamento é adotada pela abordagem de cunho probabilística conhecida como misturas de especialistas (MEs)(LIMA; COELHO; ZUBEN, 2007; JACOBS et al., 1991), que emprega um módulo de combinação especial, chamado de gating, o qual, por sua vez, faz uso da informação de entrada de maneira a selecionar os especialistas mais aptos para tratarem a entrada corrente. Recentemente, arquiteturas de MEs vêm sendo exploradas com sucesso no âmbito de classificação de sinais biomédicos (GÜLER; ÜBEYLI, 2005; SUBASI, 2007).

A outra classe de abordagens de comitês, conhecida por ensembles (coletâneas) de estimadores (HASHEM, 1997; KUNCHEVA, 2004; SHARKEY, 2012), segue a lógica de se fundir apropriadamente o conhecimento adquirido por vários módulos-especialistas para se chegar a uma decisão geral que seja supostamente superior àquela alcançável por qualquer um dos membros do grupo em separado. Neste caso, em vez de dividir a tarefa de predição em subtarefas, adota-se a redundância de diferentes módulos, os quais devem apresentar alta diversidade de generalização entre si. Os vários esquemas possíveis de combinação existentes na literatura diferem basicamente entre si com relação às suas arquiteturas, às características de seus módulos-combinadores e ao modo como produzem e selecionam seus componentes. Inclusas estão nesse contexto as proeminentes técnicas estatísticas de Bagging e Boosting (FRIEDMAN; HASTIE; TIBSHIRANI, 2001; KUNCHEVA, 2004), voltadas à geração de estimadores complementares mediante a reamostragem aleatória e adaptativa dos dados amostrais.

Costuma-se dividir o projeto efetivo de ensembles em três etapas sequenciais e interdependentes, destinadas à criação, seleção e combinação dos componentes, os quais, por sua vez, devem apresentar duas propriedades básicas: (1) elevados níveis de acuidade e (2) diferentes padrões de generalização. Nesse âmbito, ensembles heterogêneas de redes neurais feedforward, de máquinas de vetores-suporte e mesmo de diferentes indutores de

aprendizado de máquina supervisionado são abordagens introduzidas na literatura pelo proponente deste projeto (COELHO; LIMA; ZUBEN, 2003; LIMA; COELHO; ZUBEN, 2002; SALGADO et al., 2006). Pretende-se, no curso deste projeto, investigar a aplicação sistemática dessas abordagens para fins de classificação supervisionada de sinais biomédicos visando reconhecimento biométrico.

Algoritmos de comitês de máquinas também vêm sendo adotados no contexto não-supervisionado, dando origem aos modelos de ensembles de agrupamento (clustering ensembles) (HADJITODOROV; KUNCHEVA; TODOROVA, 2006; FERN; BRODLEY, 2004; TRESP, 2001; STREHL; GHOSH, 2002). A ideia principal é a de se gerar um número significativo de diferentes partições sobre os mesmos dados e, daí, combiná-las usando uma função-consenso de modo a se criar um agrupamento final capaz de incorporar todas as informações contidas nas partições-base. De maneira mais formal, considere um conjunto de amostras hipotético $D = d_1, d_2, \cdots, d_n$ e um conjunto de partições-base $P = p_1, p_2, \cdots, p_m$ obtidas pela aplicação de m algoritmos de agrupamento diferentes (ou via o mesmo algoritmo, mas com condições iniciais ou valores de parâmetros de controle diferentes). O objetivo de uma técnica de ensemble de agrupamento é, portanto, a de usar as soluções individuais em P para particionar D em grupos, de tal modo que a informação compartilhada entre as partições-base seja maximizada. Assim como no contexto supervisionado, a diversidade entre as partições-base passa a ser um ingrediente-chave para o sucesso de técnicas de ensembles de agrupamento (TOPCHY; JAIN; PUNCH, 2005; HADJITODOROV; KUNCHEVA; TODOROVA, 2006).

Vários algoritmos de ensembles de agrupamento vêm sendo propostos recentemente, tendo em vista se melhorar a precisão, estabilidade e robustez de algoritmos de agrupamento tradicionais (TOPCHY; JAIN; PUNCH, 2005). As formulações principalmente diferem em como o objetivo de maximização da informação de compartilhamento é escolhido. Um atributo comum entre os diferentes algoritmos de ensembles de agrupamento é que eles geralmente empregam uma construção gráfica como primeiro passo. Elementos em pares (pares de clusters ou pares de itens) são, então, avaliados e a seus vértices são atribuídos pesos que refletem sua similaridade. Essas abordagens apresentam como pontos vantajosos escalabilidade, paralelismo e a habilidade para capturar estruturas de dados complexas e robustez a ruído (STREHL; GHOSH, 2002). Todavia, a aplicação dessas abordagens em tarefas de classificação (não-supervisionada) de sinais biomédicos é uma linha de pesquisa ainda pouco explorada.

5.2 Abordagem Proposta

Uma condição para que um ensemble de classificadores tenha capacidade de apresentar um desempenho superior a qualquer um de seus componentes individuais é fazer com que os classificadores que irão compor o ensemble tenham um bom desempenho isoladamente e sejam diversos entre si (HASHEM, 1997). Tomado isoladamente, um classificador seria considerado um candidato a compor um ensemble se apresentar, para todas as classes envolvidas, um desempenho superior àquele produzido por um classificador aleatório, ou seja, um classificador que rotula uma amostra qualquer de entrada com um índice aleatório que apresenta uma distribuição uniforme entre as $K+1$ classes candidatas. Dois classificadores são considerados diversos se eles não apresentam os mesmos erros de classificação frente a um mesmo conjunto de amostras, ou seja, se as amostras classificadas erroneamente pelos classificadores diferem em algum grau. É evidente que podem ser definidos índices que meçam a qualidade individual e a diversidade entre os classificadores, indicando o grau de divergência.

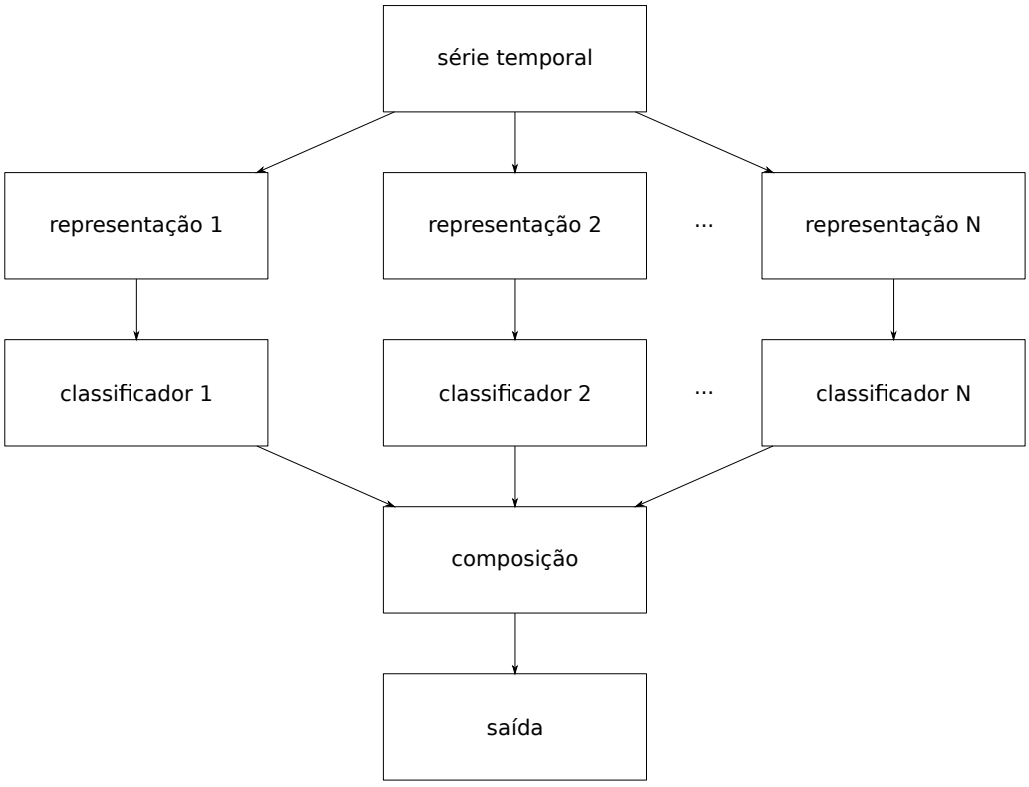
As técnicas de representação simbólicas individualmente apresentam algumas particularidades na redução de dimensionalidade, extração de informação dos segmentos, simbolização, etc, que as tornam sensíveis ao tipo de problema abordado. Além disso, as técnicas de representação simbólicas possuem parâmetros que precisam ser definidos a priori e caso não sejam especificados adequadamente pode ter uma degradação no seu desempenho. Com base nisto, pode-se dizer que as técnicas de representação simbólicas são candidatas a compor um ensemble. O ensemble de técnicas de representação simbólica, proposta desta dissertação, consiste de um conjunto de técnicas de representações simbólicas, tendo como alvo a tarefa de reconhecimento biométrico baseado em sinais de ECG. A figura 26 ilustra a abordagem proposta.

Enquanto uma arquitetura típica envolvendo uma técnica de representação simbólica opera com somente uma configuração de máquina de aprendizado, um ensemble utiliza uma combinação de M técnicas de representação simbólica, que diferem entre si, pela técnica de redução da dimensionalidade, extração de informação e simbolização, conforme descrito no Capítulo 3. Como mencionado anteriormente, um dos objetivos básicos desta iniciativa é fornecer um mecanismo automático que retira do projetista a difícil tarefa de escolher, por exemplo, a técnica de representação simbólica e seus parâmetros. Além disso,

devido ao emprego de diversas técnicas de representação simbólicas, cria-se condições de combinar múltiplas soluções alternativas, com possíveis ganhos de desempenho frente à melhor solução individual.

Por outro lado, o preço a pagar é evidente: é necessário resolver o mesmo problema tantas vezes quantos forem os candidatos a comporem o ensemble, além da necessidade de alocar recursos computacionais para realizar a seleção e combinação de modelos. Outro aspecto relevante diz respeito ao número M de componentes. Quanto maior M, menor é a possibilidade de divergência em generalização. Sendo assim, os processos de seleção normalmente alocam menos que 10 componentes para o ensemble, mesmo que existem dezenas disponíveis. Na presença de um método de seleção de modelos, o número de candidatos gerados é sempre maior ou igual ao número de componentes do ensemble. Nesta dissertação, a seleção dos componentes do ensemble é realizada via algoritmo evolutivos. Por questões de custo computacional, o número máximo de componentes também foi limitado igual a 6.

Figura 26 – Ensemble de técnicas de representação simbólica



Fonte: Henrique Passos, 2018

5.2.1 Seleção de componentes do comitê

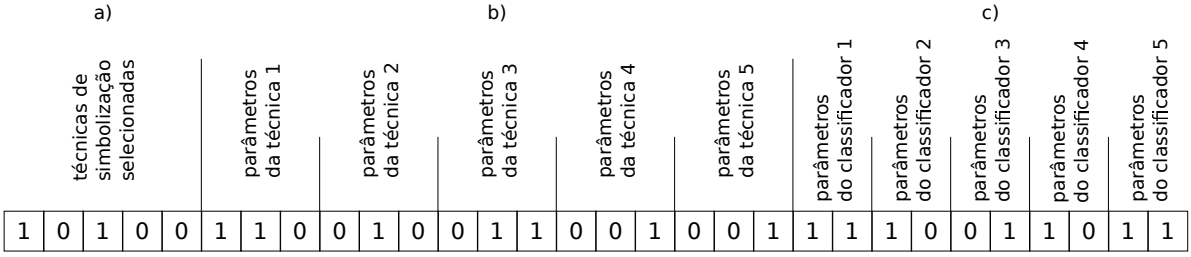
Em um comitê de classificadores, à medida que se aumenta o número de componentes, P, a suposição que todos os erros cometidos sejam mutuamente independentes não é verdadeira. Quando esta suposição falha, adicionar mais componentes ao grupo incorre em perda de recursos computacionais, uma vez que o desempenho do comitê não será melhorado. Além disso, isso pode ser prejudicial no sentido de que ao adicionar componentes com desempenho muito ruim, pode-se comprometer o desempenho do classificador resultante. Assim, seria muito importante se fosse possível encontrar o subconjunto P' indicando o grupo ótimo de componentes. Pode-se tentar todos os $2^M - 1$ possíveis subconjuntos não vazios de 2^P, mas para M grande, está busca torna-se bastante custosa. Como alternativa, (PERRONE; COOPER, 1995) propôs ordenar os componentes de acordo com o aumento dos valores de erro quadrático médio (EQM) e, em seguida, gerar um conjunto de classificadores adicionando sucessivamente os elementos de M. A cada adição de um componente, é verificador se o EQM do comitê está diminuindo ou não. Em caso positivo, o componente é mantido, caso contrário, é descartado.

Conforme mencionado no Capítulo 4, Algoritmos Evolutivos constituem uma técnica de otimização global que têm mostrado ser bem-sucedidas em domínios altamente complexos e, portanto, parece ser apropriado para lidar com as questões de complexidade envolvendo o espaço de possíveis combinações de técnicas de representação simbólica. Tais questões incluem discretização, não diferenciabilidade, alta sensibilidade (pequenas variações nas combinações podem implicar em mudanças no valor da função objetivo) e múltiplas modalidades (diferentes conjuntos de técnicas de representação simbólica podem apresentar desempenho igual). Nesse contexto, (WU; CHEN, 2001; ZHOU et al., 2002) também aplicaram uma metodologia baseada em GA para encontrar um subconjunto ótimo de componentes de redes neurais. A combinação resultante alcançou desempenho superior àqueles obtidos por métodos bastante conhecidos na área de comitê de máquinas, tais como, Boosting, Arcing e Bagging.

Com base no exposto acima, a abordagem proposta emprega algoritmos evolutivos para realizar a busca pelo subconjunto ótimo de técnicas de representação e seus parâmetros. Neste caso, o cromossomo (partícula) foi codificado como string binária. A figura 27 ilustra a codificação usada nos algoritmos evolutivos, sendo: a) os primeiros genes (onde cada

bit indica a presença (1) ou ausência (0) de uma técnica de representação simbólica) dos cromossomos representam quais técnicas de representação simbólica irá compor o ensemble; b) neste os genes representam os parâmetros das técnicas de representação simbólicas selecionadas em a); c) os genes representam os parâmetros do classificador, neste caso foi adotado o classificador K-Vizinhos Mais Próximos e este codifica o valor desse K. A

Figura 27 – Representação dos genes no GA

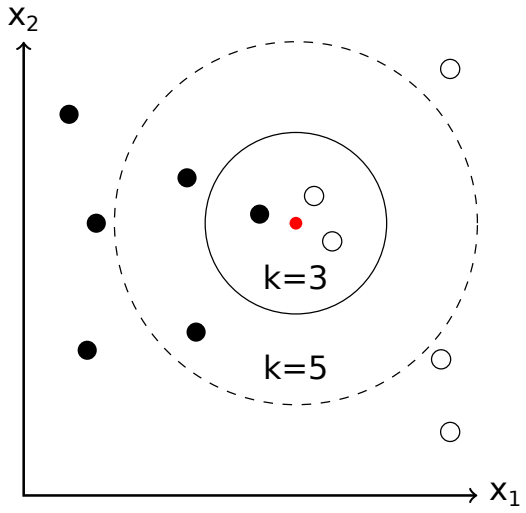


Fonte: Henrique Passos, 2018

5.2.2 K-Vizinhos Mais Próximos

O classificador K-Vizinhos Mais Próximos (do inglês K-Nearest Neighbors - KNN) (COVER; HART, 1967) pertence à família de técnicas de aprendizado baseado em instância (do inglês instance-based learning) (AHA; KIBLER; ALBERT, 1991), visto que as instâncias do conjunto de treinamento são utilizadas para prever as instâncias do conjunto de teste. Neste caso, para prever uma dada instância do conjunto de teste é realizado o voto majoritário das K instâncias do conjunto de treinamento mais próximas desta, onde K um parâmetro a ser definido pelo usuário, como ilustrado na figura 28. Um caso particular do classificador KNN é o 1-Vizinhos Mais Próximos (do inglês 1-Nearest Neighbors - 1NN), que assume o valor de K igual a um. O classificador KNN é muito utilizado em biometria (AGRAFIOTI; HATZINAKOS, 2008; SHEN; TOMPKINS; HU, 2010) e em mineração de séries temporais (KEOGH et al., 2001a; XING; PEI; KEOGH, 2010). Nesta dissertação foi adotado o classificador KNN por ser um dos classificadores mais empregados para reconhecimento de batimentos cardíacos (ODINAKA et al., 2012). Além disso, foi adotado a distância euclidiana por ser amplamente utilizada em mineração de séries temporais (KEOGH et al., 2001a).

Figura 28 – Classificador KNN



Fonte: Henrique Passos, 2018

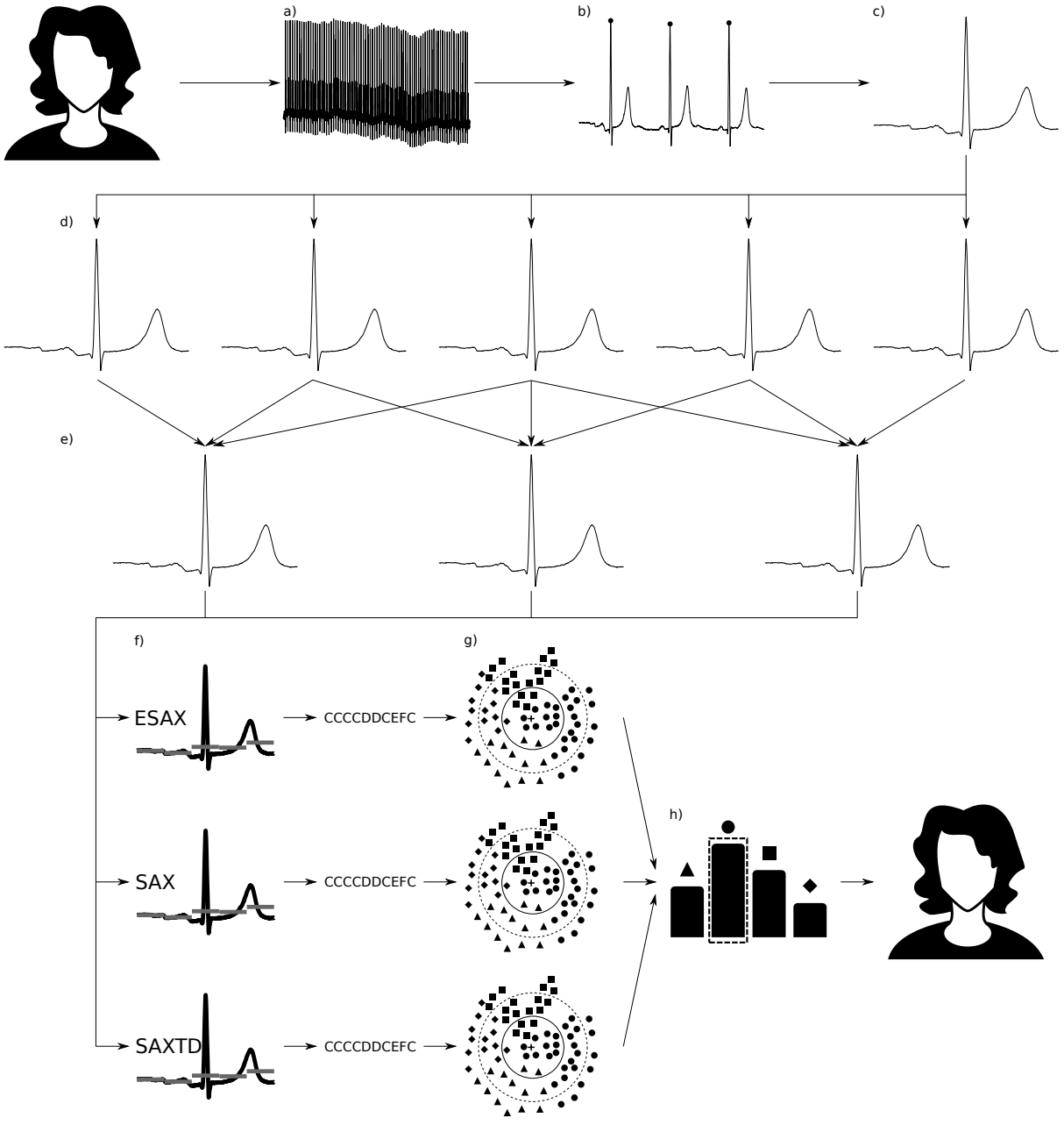
5.2.3 Extração de características

Na literatura, diversos métodos têm sido propostos para tratar o sinal de ECG como um candidato a biometria para identificação da identidade (IRVINE et al., 2008; IRVINE; ISRAEL, 2009; SHEN; TOMPKINS; HU, 2002; SINGH; GUPTA, 2011; SINGH; GUPTA, 2008; WANG et al., 2007; SINGH; GUPTA, 2009). Estes trabalhos podem ser categorizados como dependente ou independente das características fiduciais. Pontos fiduciais são aqueles pontos de interesse em um batimento cardíaco, tais como, os picos das ondas P, Q, R, S e T. Abordagens baseadas em pontos fiduciais analisam e extraem características locais para projeto de um sistema biométrico, tais como, diferenças temporais ou amplitudes entre pontos fiduciais consecutivos (SINGH; GUPTA, 2011; SINGH; GUPTA, 2008; ISRAEL et al., 2005). Por outro lado, abordagens independentes dos pontos fiduciais tratam o sinal de ECG ou o batimento cardíaco isolado holisticamente e extrai característica baseada na morfologia da forma de onda (IRVINE et al., 2008; IRVINE; ISRAEL, 2009). Esta distinção tem uma analogia para sistemas biométricos baseado em face, onde a primeira abordagem (dependente dos pontos fiduciais) opera localmente e extrai característica tais como distância entre os olhos ou o tamanho da boca. Uma abordagem holística deverá analisar a imagem facial globalmente. Alguns trabalhos utilizam ambas as abordagens para extração de características (SHEN; TOMPKINS; HU, 2002; WANG et al., 2007).

Com base no exposto acima, a abordagem proposta pode categorizada como baseada em características não fiduciais. A figura 29 ilustra como as representações simbólicas

obtidas a partir do sinal de ECG foram aplicadas no reconhecimento biométrico. Neste caso, inicialmente é realizada a remoção do baseline, isto é, interferências provenientes de algum tipo de artefato. Em seguida, é realizada a segmentação dos batimentos cardíacos. Os batimentos cardíacos são normalizados com média zero e desvio padrão um, isso evita erros no cálculo de similaridade. Após a normalização dos batimentos cardíacos foi realizada uma suavização destes por meio de um filtro do tipo média móvel, cujo objetivo era reduzir o nível de ruído. Então as técnicas de representação simbólicas são aplicadas a cada batimento extraindo uma string que será utilizada como entrada para classificador KNN.

Figura 29 – Processo de classificação: a) remoção do baseline, b) detecção do complexo QRS, c) sinal de ECG normalizado, d) batimentos cardíacos truncados, e) média de batimentos cardíacos, f) representação de séries temporais (ensemble), g) realizar a comparação entra as representações (ensemble), h) selecionar através do voto majoritário (ensemble)



Fonte: Henrique Passos, 2018

6 Resultados

Esse capítulo descreve os resultados obtidos pela abordagem proposta, cujo o objetivo é explorar conjuntamente algumas potencialidades advindas das técnicas de representação simbólica com aquelas da abordagem ensemble. Além disso, informações sobre o conjunto de dados, detalhes de pré-processamento e parametrização são descritos em detalhes.

6.1 Conjunto de dados

PhysioNet (GOLDBERGER et al., 2000) disponibiliza várias bases de dados contendo sinais fisiológicas e software para processamento dessas bases. As bases de dados de ECG apresentam informações relevantes relativas aos indivíduos e contempla uma diversidade de indivíduos em relação a sexo, idade e patologias cardíacas. Dentre as bases de dados, destaca-se a base de dados chamada PTB (do alemão: Physikalisch Technische Bundesanstalt in Braunschweig - Instituto Federal Físico e Técnico em Braunschweig) (BOUSSELJOT; KREISELER; SCHNABEL, 1995; KREISELER; BOUSSELIOT, 1995) que foi utilizada para validar a abordagem proposta neste trabalho. A base de dados PTB é composta por 549 gravações de 290 indivíduos com idades entre 17 a 87 anos. A média de idade dentre os 209 indivíduos do sexo masculino é de 57.2 anos e a média de idade dentre os 81 indivíduos do sexo feminino é de 55.5 anos. Há algumas informações faltantes sobre os indivíduos no conjunto de dados, como a idades de 1 indivíduo do sexo feminino e 14 indivíduos do sexo masculino. Dentre os 290 indivíduos, apenas 52 indivíduos são saudáveis, que corresponde a pouco menos de um quinto. Na tabela 2 é apresentado de forma sumarizada os tipos doenças cardíacas acometidas pelos indivíduos presentes no PTB. O sinal de ECG coletado consiste de:

- 16 canais de entrada, (14 de ECGs, 1 de respiração, 1 de tensão da linha);
- Tensão de entrada: $\pm$ 16 mV, compensando tensão de até $\pm$ 300 mV;
- Resistência de entrada: 100 Ω (DC);
- Resolução: 16 bit com 0.5 μV/LSB (2000 A/D unidades por mV);
- Largura da banda: 0 - 1 kHz (sincronismo de amostragem de todos os canais);

- Tensão de ruído: max. 10 μV (pp), respectivamente 3 μV (RMS) com entrada de curto-circuito;
- Gravação em linha de resistência da pele;
- Gravação de nível de ruído durante a coleta de sinal.

Tabela 2 – Diagnostico dos indivíduos do conjunto de dados do PTB

Patologia	Número de indivíduos
Infarto do miocárdio	148
Miocardiopatia/Insuficiência cardíaca	18
Bloqueio átrio-ventricular do primeiro grau	15
Arritmia cardíaca	14
Hipertrofia do miocárdio	7
Doença da válvula cardíaca	6
Miocardite	4
Variados	4
Saudáveis	52

Fonte: Henrique Passos, 2018

Cada indivíduo tem de 1 a 5 gravações, sendo que cada gravação é composta por 15 sinais coletados simultaneamente, sendo 12 derivações convencionais (i, ii, iii, avr, avl, avf, v1, v2, v3, v4, v5, v6) e 3 derivações de Frank (vx, vy, vz). Cada sinal foi digitalizado a 1000 amostras por segundo (1 kHz), com 16 bit de resolução sobre um intervalo de $\pm$ 16.384 mV. De forma geral, o PTB é uma base de dados com grande diversidade de indivíduos e tem boa qualidade nos sinais devido a alta taxa de amostragem de 1 kHz. A grande diversidade do conjunto de dados, com indivíduos normais e indivíduos com alguma patologia cardíaca, proporciona uma avaliação mais adequada da abordagem proposta.

A segunda base de dados (*MIT-BIH Arrhythmia Database*) contém o sinal de ECG obtido de 47 pessoas amostrados a $360Hz$, com duração aproximada de 30 minutos. Os registros foram escolhidos em um conjunto contendo mais de 4000 gravações obtidos a partir de (GOLDBERGER et al., 2000) *Holters* do *Beth Israel Hospital Arrhythmia Laboratory*, entre os anos de 1975 e 1979.

6.2 Validação cruzada k-fold

A base de dados foi particionada em duas partes: a) conjunto de treinamento, o qual foi destinado exclusivamente ao aprendizado do classificador; b) conjunto de avaliação para estimar a capacidade de generalização do classificador. Conforme mencionado anteriormente,

foi utilizado algoritmo evolutivo para realizar a busca pelo subconjunto ótimo de técnicas de representação simbólica. Ao longo do processo de evolução, torna-se necessário avaliar cada solução proposta por um indivíduo ou partícula na população. Após o processo evolutivo, a melhor solução encontrada deve ser avaliada sobre um outro conjunto. De forma a realizar uma avaliação adequada ao longo do processo evolutivo e da melhor solução alcançada, foi aplicado validação cruzada k-fold sobre o conjunto de avaliação. Neste caso, o conjunto de avaliação foi dividido em k partes, $k-1$ partes foi utilizada para avaliar a solução proposta por cada indivíduo ao longo do processo evolutivo e a parte restante foi utilizada para avaliar a melhor solução encontrada. Este procedimento foi repetido k vezes, desta forma garantimos que todo o conjunto de avaliação fosse utilizado para teste. O resultado apresentado consiste da média e o desvio padrão alcançado pelo melhor individuo nas k repetições.

6.2.1 Teste de Wilcoxon

O teste de Wilcoxon é um procedimento estatístico não paramétrico que ajuda a responder a seguinte pergunta: duas amostras independentes, digamos x e y, representam duas populações diferentes? A hipótese nula é que os dados em x e y são amostras de distribuições contínuas com medianas iguais. Supondo um nível de significância igual a 5%, um valor de p menor que 0.05, indica que o teste de Wilcoxon rejeita a hipótese nula, logo a diferença de desempenho entre o ensemble de representações simbólicas e os seus componentes são estatisticamente significativas (HOLLANDER; WOLFE; CHICKEN, 2013). No nosso caso, a amostra x sempre corresponde ao desempenho alcançado pelo ensemble de representação simbólica e a amostra y os componentes do Ensemble.

6.2.2 Taxa de erro igual

Um dos aspectos de maior importância nos sistemas biométricos é o seu desempenho, sendo um fator representante da segurança para o reconhecimento biométrico. Na biometria a taxa de falsa aceitação (do inglês False Acceptance Rate - FAR), a taxa de falsa rejeição (do inglês False Rejection Rate - FRR) e a taxa de erro igual (do inglês Equal Error Rate - EER) são presentes em diversos trabalhos como uma métrica para avaliar o desempenho do

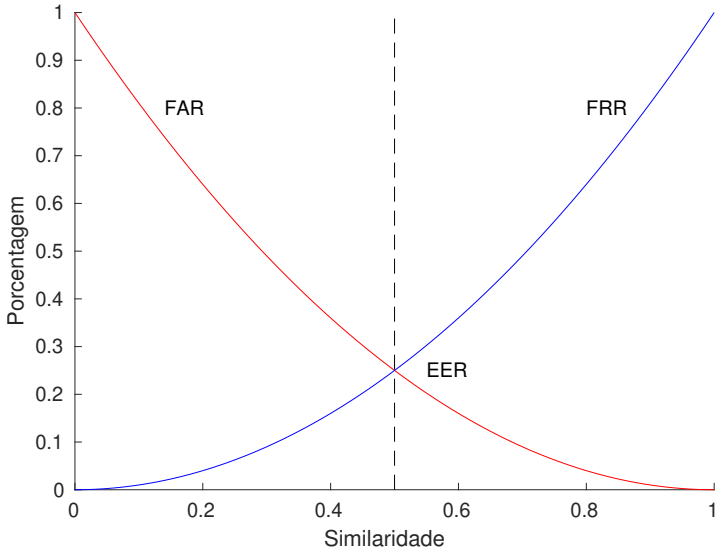
sistema biométrico (ROSS; JAIN, 2003; WANG; TAN; JAIN, 2003), sendo fundamentais para avaliar o comportamento da aceitação/rejeição de indivíduos pelo sistema biométrico e servem como base para devidos ajustes de sensibilidade do reconhecimento biométrico.

FAR representa o erro do sistema biométrico ao aceitar indivíduos que não estão autorizados e FRR está relacionado a rejeição de indivíduos devidamente autorizado pelo sistema biométrico. A taxa de falsa aceitação e rejeição estão ligadas ao limiar de similaridade, o resultado do reconhecimento biométrico é a identidade do indivíduo reconhecido e o grau de similaridade comparado com características presentes no banco de dados do sistema biométrico. Nesse trabalho essa similaridade é calculada com base na equação 10.

$$\text{similaridade} = \frac{(\text{características do individuo} - \text{características do banco de dados})^2}{\text{quantidade de características}} \quad (10)$$

A figura 30 ilustra as taxas empregues na análise do desempenho do sistema biométrico, é possível notar que o EER é o resultado do encontro de FAR com FRR. O eixo de similaridade mostra a ação desse limiar sobre a taxa de falsa aceitação e rejeição, o ponto de igualdade entre FAR e FRR é chamado de taxa de igualdade de erro. Nesse trabalho, a taxa de erro igual (EER) define o limiar de similaridade para avaliar a desempenho do Ensemble de técnicas de representação simbólica para reconhecimento biométrico baseado em sinais de ECG no conjunto de teste (seção 6.4).

Figura 30 – Ilustração da taxa de falsa aceitação (FAR), taxa de falsa rejeição (FFR) e taxa de erro igual (EER)



Fonte: Henrique Passos, 2018

6.3 Experimentos Preliminares

O objetivo deste experimento é avaliar o potencial isolado de cada técnica de representação simbólica para reconhecimento biométrico baseado em sinais de ECG. Neste experimento, foi definido um intervalo para os parâmetros de cada técnica, para cada valor dentro deste intervalo foi utilizado validação cruzada k-fold, com k igual a 10. Os melhores resultados alcançados por cada técnica em termos de taxa de reconhecimento (média e desvio padrão) para as bases de dados PTB e MIT são apresentados nas tabelas 3, 4 respectivamente. Considerando a técnica que alcançou maior taxa média de reconhecimento como amostra x, foi utilizado o teste de Wilcoxon havia diferença significativa entre as técnicas. A tabela 5 apresenta os resultados obtidos usando este teste estatístico.

No total foram realizadas 1596 execuções envolvendo a base de dados do MIT e PTB. Em mais de 56% das execuções foi possível alcançar uma taxa média de reconhecimento maior que $95,2\%$. O desvio padrão para todas as execuções se mantiveram baixo. Isto demonstra a capacidade de generalização das técnicas de representação simbólica. O DESAX e GASAX não obtiveram bons resultados quando comparado com as outras técnicas. Dentre os possíveis motivos para a baixa performance, pode-se destacar: i) a função objetivo adotada por estes algoritmos pode não ter expressado o problema de forma satisfatória no processo de otimização; ii) quantidade de épocas baixo e tamanho da população inadequados para o problema tratado.

Vale a pena salientar que os parâmetros adotados no DESAX e GASAX correspondem aos valores definidos pelos autores. O SAX-Kmeans e ESAX-Kmeans foram as técnicas que obtiveram melhores resultados para ambas as bases de dados. Essas duas técnicas têm em comum a busca pelo melhor ponto de quebra para definição das regiões a serem simbolizadas. As representações que obtiveram os menores tempos de execução foram: APCA, EFD, EFVD, ESAX, EWD, SAX, SAXTD e SPTA. Em função disso, essas técnicas foram escolhidas como candidatas para compor o ensemble. Na próxima seção, os resultados alcançados com o ensemble usando essas técnicas como candidatas são apresentados.

Tabela 3 – Taxa de reconhecimento obtida por cada técnica de representação simbólica usando a base de dados PTB

Repr.	Média ± desvio padrão	Parâmetro
1DSAX	96.293126 ± 0.003344	$\lambda = 60; \mu = 3; \alpha = 60; \nu = 2; \kappa = 5$
APCA	97.766861 ± 0.003009	$\lambda = 60; \mu = 3; \tau = 60; \kappa = 5$
ASAX	97.785576 ± 0.002287	$\lambda = 80; \mu = 3; \alpha = 60; \nu = 4; \kappa = 5$
DESAX	00.423191 ± 0.000156	$\lambda = 80; \mu = 2; \alpha = 20; \nu = 2; \kappa = 5$
DWT	97.914255 ± 0.002233	$\lambda = 80; \mu = 3; \omega = 2; \kappa = 5$
EFD	98.076614 ± 0.001243	$\lambda = 80; \mu = 3; \alpha = 60; \nu = 4; \kappa = 5$
EFVD	98.174906 ± 0.001267	$\lambda = 80; \mu = 3; \alpha = 60; \nu = 4; \kappa = 5$
ENSAX	97.592762 ± 0.002518	$\lambda = 80; \mu = 3; \alpha = 60; \nu = 4; \kappa = 5$
ESAX	97.812104 ± 0.002701	$\lambda = 80; \mu = 3; \alpha = 60; \nu = 4; \kappa = 5$
ESAX-Kmeans	98.817185 ± 0.001901	$\lambda = 80; \mu = 3; \alpha = 20; \nu = 2; \kappa = 5$
EWD	98.216465 ± 0.003611	$\lambda = 60; \mu = 3; \alpha = 40; \nu = 2; \kappa = 5$
GASAX	48.834031 ± 0.005765	$\lambda = 60; \mu = 2; \alpha = 60; \nu = 4; \kappa = 15$
R-Kmeans	98.805992 ± 0.002294	$\lambda = 60; \mu = 3; \nu = 2; \kappa = 5$
PAA	97.861349 ± 0.002237	$\lambda = 80; \mu = 3; \nu = 2; \kappa = 5$
RSAX	97.793226 ± 0.002516	$\lambda = 80; \mu = 3; \alpha = 60; \nu = 4; \phi = 4; \kappa = 5$
SAX	97.732801 ± 0.002259	$\lambda = 80; \mu = 3; \alpha = 60; \nu = 2; \kappa = 5$
SAX-Kmeans	98.821145 ± 0.001625	$\lambda = 80; \mu = 3; \alpha = 20; \nu = 4; \kappa = 5$
SAXTD	97.778070 ± 0.003078	$\lambda = 80; \mu = 3; \alpha = 60; \nu = 4; \kappa = 5$
SPTA	66.764011 ± 0.009679	$\lambda = 80; \mu = 3; \alpha = 60; \nu = 2; \kappa = 5$
VWSAX	97.729021 ± 0.002276	$\lambda = 80; \mu = 3; \alpha = 60; \nu = 2; \chi = 1.5; \kappa = 5$

Fonte: Henrique Passos, 2018

Tabela 4 – Taxa de reconhecimento obtida por cada técnica de representação simbólica usando a base de dados MIT

Repr.	Média ± desvio padrão	Parâmetro
1DSAX	93.842571 ± 0.002963	$\lambda = 80; \mu = 3; \alpha = 60; \nu = 2; \kappa = 5$
APCA	97.252116 ± 0.002692	$\lambda = 60; \mu = 3; \tau = 60; \kappa = 5$
ASAX	97.249557 ± 0.002911	$\lambda = 80; \mu = 3; \alpha = 40; \nu = 2; \kappa = 5$
DESAX	02.027311 ± 0.000061	$\lambda = 80; \mu = 2; \alpha = 20; \nu = 2; \kappa = 5$
DWT	97.465859 ± 0.002936	$\lambda = 80; \mu = 3; \omega = 2; \kappa = 5$
EFD	98.167242 ± 0.001462	$\lambda = 80; \mu = 3; \alpha = 60; \nu = 2; \kappa = 5$
EFVD	98.076367 ± 0.001889	$\lambda = 80; \mu = 3; \alpha = 60; \nu = 2; \kappa = 5$
ENSAX	97.127962 ± 0.002803	$\lambda = 80; \mu = 3; \alpha = 60; \nu = 2; \kappa = 5$
ESAX	97.208601 ± 0.002606	$\lambda = 80; \mu = 3; \alpha = 60; \nu = 4; \kappa = 5$
ESAX-Kmeans	98.386094 ± 0.001973	$\lambda = 80; \mu = 3; \alpha = 20; \nu = 4; \kappa = 5$
EWD	95.947945 ± 0.002385	$\lambda = 80; \mu = 3; \alpha = 60; \nu = 2; \kappa = 5$
GASAX	76.635983 ± 0.003245	$\lambda = 80; \mu = 3; \alpha = 60; \nu = 2; \kappa = 15$
R-Kmeans	97.994413 ± 0.001620	$\lambda = 80; \mu = 3; \nu = 4; \kappa = 5$
PAA	97.619445 ± 0.002930	$\lambda = 80; \mu = 3; \nu = 2; \kappa = 5$
RSAX	97.194514 ± 0.002972	$\lambda = 80; \mu = 3; \alpha = 60; \nu = 2; \phi = 8; \kappa = 5$
SAX	97.176605 ± 0.003010	$\lambda = 80; \mu = 3; \alpha = 60; \nu = 2; \kappa = 5$
SAX-Kmeans	98.278583 ± 0.002072	$\lambda = 80; \mu = 3; \alpha = 20; \nu = 2; \kappa = 5$
SAXTD	97.134354 ± 0.003004	$\lambda = 80; \mu = 3; \alpha = 40; \nu = 2; \kappa = 5$
SPTA	65.297612 ± 0.004897	$\lambda = 60; \mu = 3; \alpha = 60; \nu = 2; \kappa = 5$
VWSAX	97.179165 ± 0.003023	$\lambda = 80; \mu = 3; \alpha = 60; \nu = 2; \chi = 1.5; \kappa = 5$

Fonte: Henrique Passos, 2018

Tabela 5 – Resultado do teste de Wilcoxon sobre o erro de validação cruzada k-fold

	PTBDB			MITDB		
Repr.	ESAX-Kmeans	R-Kmeans	SAX-Kmeans	ESAX-Kmeans	R-Kmeans	SAX-Kmeans
1DSAX	T - 0.0002	T - 0.0002	T - 0.0002	T - 0.0002	T - 0.0002	T - 0.0002
APCA	T - 0.0002	T - 0.0002	T - 0.0002	T - 0.0002	T - 0.0002	T - 0.0002
ASAX	T - 0.0002	T - 0.0002	T - 0.0002	T - 0.0002	T - 0.0002	T - 0.0002
DESAX	T - 0.0002	T - 0.0002	T - 0.0002	T - 0.0002	T - 0.0002	T - 0.0002
DWT	T - 0.0002	T - 0.0002	T - 0.0002	T - 0.0002	T - 0.0013	T - 0.0002
EFD	T - 0.0002	T - 0.0002	T - 0.0002	T - 0.0002	T - 0.0257	T - 0.0006
EFVD	T - 0.0002	T - 0.0002	T - 0.0002	T - 0.0002	T - 0.0073	T - 0.0004
ENSAX	T - 0.0002	T - 0.0002	T - 0.0002	T - 0.0002	T - 0.0002	T - 0.0002
ESAX	T - 0.0002	T - 0.0002	T - 0.0002	T - 0.0002	T - 0.0002	T - 0.0002
ESAX-Kmeans	-	F - 1.0000	F - 0.8205	-	T - 0.0010	F - 0.3256
EWD	T - 0.0004	T - 0.0013	T - 0.0006	T - 0.0002	T - 0.0002	T - 0.0002
GASAX	T - 0.0002	T - 0.0002	T - 0.0002	T - 0.0002	T - 0.0002	T - 0.0002
R-Kmeans	F - 1.0000	-	F - 0.8798	T - 0.0010	-	T - 0.0058
PAA	T - 0.0002	T - 0.0002	T - 0.0002	T - 0.0002	T - 0.0028	T - 0.0003
RSAX	T - 0.0002	T - 0.0002	T - 0.0002	T - 0.0002	T - 0.0002	T - 0.0002
SAX	T - 0.0002	T - 0.0002	T - 0.0002	T - 0.0002	T - 0.0002	T - 0.0002
SAX-Kmeans	F - 0.8205	F - 0.8798	-	F - 0.3256	T - 0.0058	-
SAXTD	T - 0.0002	T - 0.0002	T - 0.0002	T - 0.0002	T - 0.0002	T - 0.0002
SPTA	T - 0.0002	T - 0.0002	T - 0.0002	T - 0.0002	T - 0.0002	T - 0.0002
VWSAX	T - 0.0002	T - 0.0002	T - 0.0002	T - 0.0002	T - 0.0002	T - 0.0002

Fonte: Henrique Passos, 2018

6.4 Experimentos

Neste experimento, a base de dados foi dividida aleatoriamente em 70% para conjunto de treinamento e 30% para o conjunto avaliação, no qual foi utilizada validação cruzada k-fold, com k igual a 10. O número máximo de componentes no ensemble foi definido igual a 5. Cada algoritmo evolutivo foi executado 5 vezes. As técnicas de representação simbólica candidatas a compor o ensemble foram as seguintes: APCA, EFD, EFVD, ESAX, EWD, SAX, SAXTD e SPTA. O intervalo de busca adotado para cada parâmetro empregado nas técnicas de representação simbólica é descrito a seguir:

- APCA, EFD, EFVD, ESAX, EWD, SAX, SAXTD e SPTA - *janela do QRS* (λ): $[10, 128]$, *janela média* (μ): $[1, 4]$
- EFD, EFVD, ESAX, EWD, SAX, SAXTD e SPTA - *tamanho dos segmentos* (ν): $[1, 10]$
- EFD, EFVD, ESAX, EWD, SAX, SAXTD e SPTA - *tamanho do alfabeto* (α): $[3, 64]$
- APCA - *quantidade de segmentos* (τ): $[1, 4]$

Em relação aos algoritmos evolutivos, a função objetivo é fundamental para os algoritmos evolutivos, ela avalia o desempenho dos indivíduos da população a um problema.

Na abordagem proposta a função objetivo é baseado no reconhecimento dos batimentos cardíacos através do ensemble de técnicas de representação simbólica, no reconhecimento a validação cruzada é aplicada para avalia a capacidade de generalização do reconhecimento dos batimentos cardíacos. O resultado da função objetivo é o erro do reconhecimento dos batimentos cardíacos através do ensemble, o erro assume valores de zero a um, sendo um o erro máximo e zero o erro mínimo do reconhecimento.

Na abordagem proposta todos os algoritmos adotam o tamanho da população igual a cem. A população está sucessível a operações que podem afetar até mesmo o seu melhor individuo, o elitismo transfere a cópia do melhor individuo para a próxima iteração (geração). Em relação aos critérios de parada para os algoritmos apresentados nesse capitulo existem três critérios de parada que foram definidos: o número máximo de iterações igual a quinhentos ou o erro do reconhecimento dos batimentos cardíacos igual a zero ou o erro do reconhecimento dos batimentos cardíacos do melhor individuo estável por cinquenta gerações. Qualquer um desses critérios enceram as iterações.

No algoritmo genético com população múltipla na seleção dos indivíduos foi adotado a amostragem estocástica uniforme com a taxa de seleção de 80% para as subpopulações, dentre os indivíduos selecionados 100% dos indivíduos são recombinados e 2.5% dos genes da população seleciona sofrem mutações. Dentro os indivíduos selecionados 90% voltam a serem reinseridos na população para a próxima iteração. Os cem indivíduos são distribuídos em cinco subpopulações, o processo de migração é baseado na topologia de vizinhança e a cada vinte gerações 20% dos indivíduos são migrados.

No algoritmo de evolução diferencial para potencializar a exploração do espaço de busca o fator de mutação (F) adotado foi de 80%. A probabilidade de recombinação (CR) no DE é de 50%. No PSO com fator de enxugamento (CPSO) e com codificação binaria (BCPSO) o próprio parâmetro fator de enxugamento χ assume o valor de 0.72984. φ_1 e φ_2 assumem o valor de 1.5500. O $\phi < 4$ proporciona a busca de soluções ao redor do melhor valor.

Analisando os resultados, as tabelas 6 e 7 apresentam os resultados alcançados pelo ensemble com os componentes selecionados pelo BCPSO para as bases de dados PTB e MIT respectivamente. A taxa (tx) apresentada nas tabelas a seguir corresponde a taxa do reconhecimento dos batimentos cardíacos, sendo a acurácia da classificação dos batimentos cardíacos pelo ensemble. A taxa de treinamento é referente ao melhor resultado obtido pelo algoritmo evolutivo na construção do ensemble e a taxa de teste

é correspondente ao reconhecimento dos batimentos cardíacos do ensemble no conjunto de teste. As tabelas também apresentam duas construções de ensemble, por otimização global e individual. A otimização global emprega os algoritmos evolutivos na busca das técnicas de representação simbólica e seus parâmetros simultaneamente, já a otimização local assume as técnicas de representação simbólica selecionadas pela otimização global e otimizar apenas os parâmetros das técnicas individualmente. Dentre as duas formas de construção dos ensembles, esse trabalho foca na otimização global, pelo fato da construção do ensemble ser mais dinâmica e ser favorecido pela geração de componentes com maior potencial de diversidade.

Tabela 6 – Resultado obtido pelo Ensemble com BCPSO usando a base de dados PTB

Resultados do Ensemble por otimização global					
Repr.	Tx. treino	Tx. teste	EER	Wilcoxon	Parâmetro
EFD	99.1687	99.0839	69.1219	V - 0.0002	$\lambda = 110; \mu = 4; \alpha = 29; \nu = 4; \kappa = 2$
EFVD	98.0843	97.6312	77.9495	V - 0.0002	$\lambda = 103; \mu = 1; \alpha = 62; \nu = 8; \kappa = 2$
ESAX	96.3916	92.8657	87.3971	V - 0.0002	$\lambda = 50; \mu = 4; \alpha = 46; \nu = 9; \kappa = 21$
EWD	99.8753	99.7502	93.8373	F - 0.2247	$\lambda = 114; \mu = 4; \alpha = 36; \nu = 1; \kappa = 1$
SAXTD	99.0176	99.3060	88.7203	V - 0.0003	$\lambda = 125; \mu = 4; \alpha = 47; \nu = 2; \kappa = 4$
Ensemble	99.8866	99.8334	96.1969	-	Representações acima

Resultados do Ensemble por otimização individual					
Repr.	Tx. treino	Tx. teste	EER	Wilcoxon	Parâmetro
EFD	99.7355	99.5558	44.0178	V - 0.0004	$\lambda = 104; \mu = 4; \alpha = 64; \nu = 5; \kappa = 2$
EFVD	99.7280	99.7779	79.0321	V - 0.0237	$\lambda = 122; \mu = 4; \alpha = 54; \nu = 5; \kappa = 2$
ESAX	99.5126	99.6761	82.4558	V - 0.0006	$\lambda = 123; \mu = 4; \alpha = 62; \nu = 5; \kappa = 1$
EWD	99.9622	99.9260	95.4659	F - 1.0000	$\lambda = 125; \mu = 4; \alpha = 4; \nu = 1; \kappa = 1$
SAXTD	99.5126	99.6854	85.0097	V - 0.0020	$\lambda = 116; \mu = 4; \alpha = 57; \nu = 1; \kappa = 1$
Ensemble	99.9358	99.9260	95.7065	-	Representações acima

Fonte: Henrique Passos, 2018

As tabelas 8 e 9 apresentam os resultados alcançados pelo ensemble com os componentes selecionados pelo CPSO para as bases de dados PTB e MIT respectivamente.

As tabelas 10 e 11 apresentam os resultados alcançados pelo ensemble com os componentes selecionados pelo DE para as bases de dados PTB e MIT respectivamente.

As tabelas 12, 14, 16, 18 e 20 apresentam os resultados alcançados pelo ensemble (em 5 execuções) com os componentes selecionados pelo MPGA para a base de dados PTB. Já as tabelas 13, 15, 17, 19 e 21 apresentam os resultados alcançados pelo ensemble (em 5 execuções) com os componentes selecionados pelo MPGA para a base de dados MIT.

Com relação ao classificador KNN, para a maioria dos ensembles gerados pelos algoritmos evolutivos, o parâmetro κ convergiu para o valor igual a 1. Isto significa que o batimento cardíaco do conjunto de treinamento mais próximo do batimento cardíaco do

Tabela 7 – Resultado obtido pelo Ensemble com BCPSO usando a base de dados MIT

Resultados do Ensemble por otimização global					
Repr.	Tx. treino	Tx. teste	EER	Wilcoxon	Parâmetro
EFD	99.7491	99.6637	85.9865	F - 0.2838	$\lambda = 126; \mu = 4; \alpha = 62; \nu = 9; \kappa = 1$
EFVD	99.4145	99.3834	90.4148	F - 0.0837	$\lambda = 118; \mu = 3; \alpha = 32; \nu = 1; \kappa = 1$
ESAX	98.9962	98.0381	92.6570	V - 0.0001	$\lambda = 93; \mu = 4; \alpha = 27; \nu = 7; \kappa = 5$
EWD	98.0134	96.8049	86.8834	V - 0.0001	$\lambda = 89; \mu = 3; \alpha = 63; \nu = 5; \kappa = 3$
SAXTD	90.3806	80.7175	75.5045	V - 0.0001	$\lambda = 113; \mu = 1; \alpha = 64; \nu = 8; \kappa = 30$
Ensemble	99.9791	99.8879	96.8049	-	Representações acima

Resultados do Ensemble por otimização individual					
Repr.	Tx. treino	Tx. teste	EER	Wilcoxon	Parâmetro
EFD	99.9791	99.9439	99.9439	F - 0.5428	$\lambda = 104; \mu = 4; \alpha = 34; \nu = 1; \kappa = 2$
EFVD	99.9582	99.8879	86.7152	F - 1.0000	$\lambda = 106; \mu = 4; \alpha = 51; \nu = 2; \kappa = 2$
ESAX	99.6863	99.6637	91.8722	F - 0.4865	$\lambda = 128; \mu = 4; \alpha = 36; \nu = 8; \kappa = 2$
EWD	99.9582	99.8318	91.7601	F - 0.7655	$\lambda = 114; \mu = 4; \alpha = 5; \nu = 1; \kappa = 2$
SAXTD	99.6654	99.6076	93.3296	F - 0.2254	$\lambda = 117; \mu = 4; \alpha = 23; \nu = 5; \kappa = 1$
Ensemble	100.0000	99.8879	99.9439	-	Representações acima

Fonte: Henrique Passos, 2018

Tabela 8 – Resultado obtido pelo Ensemble com CPSO usando a base de dados PTB

Resultados do Ensemble por otimização global					
Repr.	Tx. treino	Tx. teste	EER	Wilcoxon	Parâmetro
EFVD	99.5844	99.4170	50.3007	V - 0.0002	$\lambda = 128; \mu = 4; \alpha = 64; \nu = 8; \kappa = 1$
EWD	99.8715	99.7594	94.5776	F - 0.3245	$\lambda = 128; \mu = 4; \alpha = 44; \nu = 1; \kappa = 1$
SAX	99.4786	99.7409	89.0626	F - 0.1397	$\lambda = 127; \mu = 4; \alpha = 64; \nu = 1; \kappa = 1$
SAXTD	98.2770	97.8717	82.9833	V - 0.0002	$\lambda = 67; \mu = 1; \alpha = 41; \nu = 2; \kappa = 1$
SPTA	56.9674	49.0608	27.2601	V - 0.0002	$\lambda = 107; \mu = 4; \alpha = 64; \nu = 10; \kappa = 2$
Ensemble	99.9395	99.8519	95.6972	-	Representações acima

Resultados do Ensemble por otimização individual					
Repr.	Tx. treino	Tx. teste	EER	Wilcoxon	Parâmetro
EFVD	99.7128	99.5651	35.0791	V - 0.0003	$\lambda = 128; \mu = 4; \alpha = 56; \nu = 4; \kappa = 1$
EWD	99.9207	99.8982	90.8115	F - 0.3687	$\lambda = 56; \mu = 4; \alpha = 64; \nu = 1; \kappa = 1$
SAX	99.5050	99.7409	89.0719	V - 0.0042	$\lambda = 128; \mu = 4; \alpha = 59; \nu = 1; \kappa = 1$
SAXTD	99.5050	99.7409	89.0719	V - 0.0042	$\lambda = 128; \mu = 4; \alpha = 59; \nu = 1; \kappa = 1$
SPTA	83.2049	77.5053	53.4653	V - 0.0002	$\lambda = 61; \mu = 4; \alpha = 56; \nu = 1; \kappa = 1$
Ensemble	99.8904	99.9260	92.7177	-	Representações acima

Fonte: Henrique Passos, 2018

conjunto de teste é o mais significante para realizar o reconhecimento. Além disso, pode-se observar pouca variação entre o desempenho alcançado no conjunto de treinamento e aquele alcançado no conjunto de teste.

Analisando os parâmetros das técnicas de representação simbólica que compõem o ensemble é possível notar alguns padrões. Para o tamanho da janela do QRS (λ) e o tamanho do alfabeto (α), os resultados mostram que quanto maior o tamanho da janela e do alfabeto maior será desempenho, porém isso nem sempre é verdade. Na tabela 10 a representação EWD obteve uma taxa de reconhecimento igual a 99.9352% com o tamanho do alfabeto igual a 4. Isso significa que com apenas 4 regiões simbólicas o EWD conseguiu

Tabela 9 – Resultado obtido pelo Ensemble com CPSO usando a base de dados MIT

Resultados do Ensemble por otimização global					
Repr.	Tx. treino	Tx. teste	EER	Wilcoxon	Parâmetro
EFD	99.4354	99.2152	50.7848	V - 0.0105	$\lambda = 128; \mu = 4; \alpha = 27; \nu = 4; \kappa = 1$
EWD	99.4772	99.5516	91.9283	V - 0.0412	$\lambda = 128; \mu = 4; \alpha = 55; \nu = 10; \kappa = 1$
SAX	98.6826	98.4865	72.0852	V - 0.0001	$\lambda = 126; \mu = 4; \alpha = 3; \nu = 1; \kappa = 1$
SAXTD	94.0402	89.3498	78.0269	V - 0.0001	$\lambda = 77; \mu = 1; \alpha = 17; \nu = 1; \kappa = 14$
SPTA	35.5082	26.5695	13.9574	V - 0.0001	$\lambda = 105; \mu = 4; \alpha = 64; \nu = 10; \kappa = 1$
Ensemble	100.0000	99.8879	92.4327	-	Representações acima

Resultados do Ensemble por otimização individual					
Repr.	Tx. treino	Tx. teste	EER	Wilcoxon	Parâmetro
EFD	99.9791	99.8318	87.6121	F - 0.8033	$\lambda = 113; \mu = 4; \alpha = 54; \nu = 1; \kappa = 1$
EWD	99.9791	99.8318	92.0964	F - 0.7655	$\lambda = 119; \mu = 4; \alpha = 5; \nu = 1; \kappa = 1$
SAX	99.6654	99.6637	92.0404	F - 0.5184	$\lambda = 126; \mu = 4; \alpha = 46; \nu = 3; \kappa = 1$
SAXTD	99.6654	99.6637	91.4798	F - 0.5184	$\lambda = 128; \mu = 4; \alpha = 23; \nu = 1; \kappa = 1$
SPTA	78.4191	72.5336	53.5314	V - 0.0001	$\lambda = 31; \mu = 4; \alpha = 9; \nu = 1; \kappa = 1$
Ensemble	99.9582	99.8318	95.1233	-	Representações acima

Fonte: Henrique Passos, 2018

Tabela 10 – Resultado obtido pelo Ensemble com DE usando a base de dados PTB

Resultados do Ensemble por otimização global					
Repr.	Tx. treino	Tx. teste	EER	Wilcoxon	Parâmetro
EFVD	97.2606	52.1976	27.2786	V - 0.0002	$\lambda = 93; \mu = 3; \alpha = 45; \nu = 1; \kappa = 1$
ESAX	99.1310	99.4633	68.8905	V - 0.0002	$\lambda = 128; \mu = 3; \alpha = 46; \nu = 4; \kappa = 1$
EWD	99.8602	99.7039	93.6615	V - 0.0044	$\lambda = 110; \mu = 4; \alpha = 36; \nu = 1; \kappa = 2$
SAX	99.4106	99.6761	74.1279	V - 0.0010	$\lambda = 127; \mu = 4; \alpha = 49; \nu = 1; \kappa = 2$
SAXTD	98.3375	97.8810	80.3646	V - 0.0002	$\lambda = 128; \mu = 1; \alpha = 64; \nu = 2; \kappa = 1$
Ensemble	99.9509	99.8890	93.7263	-	Representações acima

Resultados do Ensemble por otimização individual					
Repr.	Tx. treino	Tx. teste	EER	Wilcoxon	Parâmetro
EFVD	99.8526	99.6114	47.1824	V - 0.0031	$\lambda = 123; \mu = 4; \alpha = 64; \nu = 3; \kappa = 2$
ESAX	99.5201	99.7409	89.3310	F - 0.3613	$\lambda = 128; \mu = 4; \alpha = 59; \nu = 2; \kappa = 1$
EWD	99.9622	99.9352	85.3336	F - 0.0597	$\lambda = 128; \mu = 4; \alpha = 4; \nu = 1; \kappa = 2$
SAX	99.5201	99.7039	84.1954	F - 0.2549	$\lambda = 118; \mu = 4; \alpha = 62; \nu = 1; \kappa = 1$
SAXTD	99.5126	99.7224	79.7076	F - 0.1116	$\lambda = 127; \mu = 4; \alpha = 56; \nu = 1; \kappa = 1$
Ensemble	99.6637	99.8242	92.1810	-	Representações acima

Fonte: Henrique Passos, 2018

obter um alta de taxa de reconhecimento. Esse tipo de comportamento mostra que há um padrão no comportamento dos parâmetros, porém esse comportamento não é estritamente regular. Há a necessidade de realizar buscas por esses parâmetros.

O algoritmo genético foi executado 5 vezes para avaliar a convergência da abordagem proposta, porém com populações iniciais distintas. Os resultados alcançados nestas execuções estão apresentados nas tabelas 12, 14, 16, 18 e 20. Neste caso, foi obtido os mesmos padrões de parâmetros mencionados anteriormente, os resultados da acurácia nos testes foram de 99.8242% até 99.9352%.

Tabela 11 – Resultado obtido pelo Ensemble com DE usando a base de dados MIT

Resultados do Ensemble por otimização global					
Repr.	Tx. treino	Tx. teste	EER	Wilcoxon	Parâmetro
APCA	2.9276	2.5224	2.5224	V - 0.0001	$\lambda = 128; \mu = 1; \tau = 1; \kappa = 10$
EFD	99.8954	100.0000	100.0000	F - NaN	$\lambda = 119; \mu = 4; \alpha = 64; \nu = 4; \kappa = 1$
EFVD	100.0000	99.8879	95.9081	F - 0.1681	$\lambda = 103; \mu = 4; \alpha = 32; \nu = 1; \kappa = 1$
ESAX	99.4981	99.3274	93.7220	V - 0.0007	$\lambda = 73; \mu = 4; \alpha = 64; \nu = 6; \kappa = 1$
EWD	99.8536	99.6637	83.0157	V - 0.0149	$\lambda = 89; \mu = 4; \alpha = 64; \nu = 2; \kappa = 1$
Ensemble	100.0000	100.0000	97.2534	-	Representações acima

Resultados do Ensemble por otimização individual					
Repr.	Tx. treino	Tx. teste	EER	Wilcoxon	Parâmetro
APCA	88.7913	81.7265	65.6951	V - 0.0001	$\lambda = 59; \mu = 4; \tau = 4; \kappa = 1$
EFD	100.0000	99.8318	95.7399	F - 0.7655	$\lambda = 126; \mu = 4; \alpha = 40; \nu = 1; \kappa = 1$
EFVD	99.9791	99.8318	69.1143	F - 0.9627	$\lambda = 105; \mu = 4; \alpha = 63; \nu = 1; \kappa = 1$
ESAX	99.7072	99.6637	91.6480	F - 0.7093	$\lambda = 128; \mu = 4; \alpha = 31; \nu = 7; \kappa = 2$
EWD	99.9791	99.8318	91.7040	F - 1.0000	$\lambda = 118; \mu = 4; \alpha = 5; \nu = 1; \kappa = 1$
Ensemble	100.0000	99.8318	96.8610	-	Representações acima

Fonte: Henrique Passos, 2018

Tabela 12 – Resultado obtido pelo Ensemble com MPGA usando a base de dados PTB

Resultados do Ensemble por otimização global					
Repr.	Tx. treino	Tx. teste	EER	Wilcoxon	Parâmetro
EFD	98.2166	95.9193	64.0696	V - 0.0002	$\lambda = 90; \mu = 4; \alpha = 61; \nu = 2; \kappa = 1$
ESAX	99.0365	99.2320	78.8563	V - 0.0002	$\lambda = 75; \mu = 3; \alpha = 62; \nu = 3; \kappa = 2$
EWD	99.8564	99.7409	94.3740	V - 0.0071	$\lambda = 112; \mu = 4; \alpha = 6; \nu = 1; \kappa = 1$
SAX	97.8425	97.4739	81.2807	V - 0.0002	$\lambda = 128; \mu = 1; \alpha = 64; \nu = 9; \kappa = 4$
SAXTD	98.8514	98.6768	76.9501	V - 0.0002	$\lambda = 128; \mu = 4; \alpha = 3; \nu = 4; \kappa = 2$
Ensemble	99.9811	99.9167	96.4930	-	Representações acima

Resultados do Ensemble por otimização individual					
Repr.	Tx. treino	Tx. teste	EER	Wilcoxon	Parâmetro
EFD	99.7317	99.5558	26.8345	V - 0.0051	$\lambda = 120; \mu = 4; \alpha = 64; \nu = 5; \kappa = 2$
ESAX	99.5126	99.6854	82.1134	F - 0.3840	$\lambda = 128; \mu = 4; \alpha = 64; \nu = 5; \kappa = 2$
EWD	99.9282	99.8982	90.8208	F - 0.0518	$\lambda = 56; \mu = 4; \alpha = 52; \nu = 1; \kappa = 2$
SAX	99.5126	99.7317	92.8380	F - 0.6493	$\lambda = 125; \mu = 4; \alpha = 59; \nu = 1; \kappa = 1$
SAXTD	99.5126	99.7224	79.7076	F - 0.5957	$\lambda = 127; \mu = 4; \alpha = 56; \nu = 1; \kappa = 1$
Ensemble	99.6562	99.7687	94.2630	-	Representações acima

Fonte: Henrique Passos, 2018

A cada execução do algoritmo genético era gerado uma população inicial distinta. Apesar disso, foi observado o mesmo comportamento mencionado anteriormente. Neste caso, a taxa de reconhecimento variou de 99.8242% a 99.9352%. Em duas execuções (tabelas 12 e 14)do algoritmo genético foi possível observar nitidamente que a abordagem proposta é estatisticamente superior aos seus componentes. No entanto, ao compor o ensemble com técnicas de representação simbólicas otimizadas individualmente não foi possível alcançar um desempenho superior em relação aos seus componentes. Isto mostra que se os componentes possuem alto desempenho e pouca discordância entre si, o ensemble não

Tabela 13 – Resultado obtido pelo Ensemble com MPGA usando a base de dados MIT

Resultados do Ensemble por otimização global					
Repr.	Tx. treino	Tx. teste	EER	Wilcoxon	Parâmetro
EFD	99.8954	99.8318	89.1256	F - 0.3290	$\lambda = 86; \mu = 4; \alpha = 37; \nu = 2; \kappa = 1$
EFVD	99.3308	98.8789	88.5650	V - 0.0006	$\lambda = 84; \mu = 4; \alpha = 62; \nu = 5; \kappa = 4$
ESAX	93.8310	91.8161	72.9260	V - 0.0001	$\lambda = 39; \mu = 1; \alpha = 53; \nu = 5; \kappa = 7$
EWD	99.6445	99.6637	88.3408	F - 0.1117	$\lambda = 119; \mu = 4; \alpha = 6; \nu = 4; \kappa = 2$
SAXTD	98.0343	97.5336	89.5179	V - 0.0001	$\lambda = 128; \mu = 2; \alpha = 59; \nu = 4; \kappa = 2$
Ensemble	99.9791	99.9439	95.6839	-	Representações acima

Resultados do Ensemble por otimização individual					
Repr.	Tx. treino	Tx. teste	EER	Wilcoxon	Parâmetro
EFD	99.9791	99.8879	96.3004	F - 1.0000	$\lambda = 117; \mu = 4; \alpha = 47; \nu = 1; \kappa = 1$
EFVD	99.9791	99.8879	87.4439	F - 0.9567	$\lambda = 120; \mu = 4; \alpha = 61; \nu = 1; \kappa = 1$
ESAX	99.7072	99.6637	91.6480	F - 0.4562	$\lambda = 128; \mu = 4; \alpha = 31; \nu = 7; \kappa = 2$
EWD	99.9373	99.7758	84.5291	F - 0.4865	$\lambda = 121; \mu = 4; \alpha = 58; \nu = 1; \kappa = 1$
SAXTD	99.6654	99.6076	92.3206	F - 0.2083	$\lambda = 112; \mu = 4; \alpha = 31; \nu = 3; \kappa = 1$
Ensemble	99.9791	99.8879	96.9731	-	Representações acima

Fonte: Henrique Passos, 2018

Tabela 14 – Resultado obtido pelo Ensemble com MPGA usando a base de dados PTB

Resultados do Ensemble por otimização global					
Repr.	Tx. treino	Tx. teste	EER	Wilcoxon	Parâmetro
EFVD	87.4896	81.5860	59.8501	V - 0.0001	$\lambda = 119; \mu = 4; \alpha = 37; \nu = 2; \kappa = 2$
ESAX	99.4899	99.6946	78.4677	V - 0.0068	$\lambda = 128; \mu = 4; \alpha = 63; \nu = 5; \kappa = 2$
EWD	99.8753	99.7409	93.5690	V - 0.0025	$\lambda = 112; \mu = 4; \alpha = 34; \nu = 1; \kappa = 1$
SAX	97.6498	96.7799	83.7420	V - 0.0001	$\lambda = 35; \mu = 1; \alpha = 64; \nu = 3; \kappa = 4$
SAXTD	98.7040	98.8618	73.1470	V - 0.0001	$\lambda = 125; \mu = 3; \alpha = 7; \nu = 8; \kappa = 1$
Ensemble	99.9736	99.9352	96.3727	-	Representações acima

Resultados do Ensemble por otimização individual					
Repr.	Tx. treino	Tx. teste	EER	Wilcoxon	Parâmetro
EFVD	99.6713	99.5373	46.7567	V - 0.0024	$\lambda = 128; \mu = 4; \alpha = 61; \nu = 5; \kappa = 2$
ESAX	99.5201	99.7409	89.3310	F - 0.4683	$\lambda = 128; \mu = 4; \alpha = 59; \nu = 2; \kappa = 1$
EWD	99.9282	99.8982	90.8208	F - 0.2663	$\lambda = 56; \mu = 4; \alpha = 52; \nu = 1; \kappa = 1$
SAX	99.5050	99.7409	89.0719	F - 0.3598	$\lambda = 128; \mu = 4; \alpha = 59; \nu = 1; \kappa = 1$
SAXTD	99.4861	99.7317	92.8565	F - 0.4482	$\lambda = 119; \mu = 4; \alpha = 34; \nu = 2; \kappa = 1$
Ensemble	99.6373	99.8057	94.3092	-	Representações acima

Fonte: Henrique Passos, 2018

produzirá melhor resultado. No melhor caso, o ensemble obteve uma taxa de reconhecimento igual a 99.9352%.

Como a quantidade de componentes no ensemble foi restringido a 5, não importa a quantidade de técnicas de representação simbólica, o custo computacional será sempre o mesmo. Considere uma população com 500 indivíduos, 100 gerações e com número máximo de componentes no ensemble igual a 5, neste caso, a quantidade de execuções será $500 \times 100 \times 5 = 250000$. Se houver 10 ou 100 técnicas de representação simbólicas candidatas, o custo será o mesmo. Em contrapartida, caso o número de componentes no ensemble seja alterado, o custo computacional altera drasticamente.

Tabela 15 – Resultado obtido pelo Ensemble com MPGA usando a base de dados MIT

Resultados do Ensemble por otimização global					
Repr.	Tx. treino	Tx. teste	EER	Wilcoxon	Parâmetro
EFD	99.7072	99.1592	81.1659	V - 0.0248	$\lambda = 111; \mu = 4; \alpha = 29; \nu = 4; \kappa = 3$
EWD	99.4981	99.0471	83.0717	F - 0.2501	$\lambda = 89; \mu = 3; \alpha = 44; \nu = 1; \kappa = 2$
SAX	99.1635	98.5987	88.7892	V - 0.0004	$\lambda = 65; \mu = 4; \alpha = 44; \nu = 10; \kappa = 2$
SAXTD	98.1179	97.3094	84.1368	V - 0.0002	$\lambda = 88; \mu = 2; \alpha = 53; \nu = 3; \kappa = 2$
SPTA	32.8105	26.1211	13.9013	V - 0.0002	$\lambda = 52; \mu = 4; \alpha = 46; \nu = 9; \kappa = 21$
Ensemble	99.9791	99.6637	94.6188	-	Representações acima

Resultados do Ensemble por otimização individual					
Repr.	Tx. treino	Tx. teste	EER	Wilcoxon	Parâmetro
EFD	99.9791	99.7758	75.1682	F - 1.0000	$\lambda = 106; \mu = 4; \alpha = 39; \nu = 1; \kappa = 2$
EWD	99.9373	99.7758	83.8565	F - 1.0000	$\lambda = 117; \mu = 4; \alpha = 63; \nu = 1; \kappa = 1$
SAX	99.6654	99.6637	91.4798	F - 0.8157	$\lambda = 128; \mu = 4; \alpha = 23; \nu = 1; \kappa = 2$
SAXTD	99.6654	99.6076	89.7422	F - 0.8157	$\lambda = 128; \mu = 4; \alpha = 26; \nu = 6; \kappa = 1$
SPTA	77.8335	71.6368	50.2803	V - 0.0001	$\lambda = 31; \mu = 4; \alpha = 42; \nu = 1; \kappa = 2$
Ensemble	99.9373	99.7758	93.9462	-	Representações acima

Fonte: Henrique Passos, 2018

Tabela 16 – Resultado obtido pelo Ensemble com MPGA usando a base de dados PTB

Resultados do Ensemble por otimização global					
Repr.	Tx. treino	Tx. teste	EER	Wilcoxon	Parâmetro
EFD	93.0439	92.7362	55.0569	V - 0.0002	$\lambda = 55; \mu = 4; \alpha = 62; \nu = 1; \kappa = 1$
ESAX	98.7267	98.6490	75.7379	V - 0.0002	$\lambda = 128; \mu = 2; \alpha = 24; \nu = 10; \kappa = 2$
EWD	99.8715	99.7964	94.8552	F - 0.7610	$\lambda = 128; \mu = 4; \alpha = 64; \nu = 1; \kappa = 1$
SAXTD	99.2783	99.4170	72.5826	V - 0.0004	$\lambda = 109; \mu = 4; \alpha = 35; \nu = 9; \kappa = 1$
SPTA	41.0451	33.3488	21.1900	V - 0.0002	$\lambda = 128; \mu = 1; \alpha = 31; \nu = 2; \kappa = 25$
Ensemble	99.9282	99.8242	94.2075	-	Representações acima

Resultados do Ensemble por otimização individual					
Repr.	Tx. treino	Tx. teste	EER	Wilcoxon	Parâmetro
EFD	99.8678	98.6768	59.5725	V - 0.0002	$\lambda = 108; \mu = 4; \alpha = 64; \nu = 2; \kappa = 2$
ESAX	99.5277	99.6576	69.5383	V - 0.0089	$\lambda = 127; \mu = 4; \alpha = 34; \nu = 4; \kappa = 2$
EWD	99.9282	99.8982	90.8208	F - 0.9694	$\lambda = 56; \mu = 4; \alpha = 52; \nu = 1; \kappa = 2$
SAXTD	99.5126	99.7224	79.7076	V - 0.0187	$\lambda = 127; \mu = 4; \alpha = 56; \nu = 1; \kappa = 1$
SPTA	83.9001	79.7909	57.0834	V - 0.0002	$\lambda = 76; \mu = 4; \alpha = 15; \nu = 1; \kappa = 1$
Ensemble	99.9358	99.8890	90.9503	-	Representações acima

Fonte: Henrique Passos, 2018

Visando identificar quais técnicas estavam presentes na população final dos algoritmos evolutivos, foi realizado uma contagem destas. A tabela 31 apresenta o percentual alcançado por cada técnica. As três representações mais selecionadas foram EWD, SAXTD e ESAX. Um aspecto interessante é que o EWD remove a tendência da série temporal para realizar a representação, já o SAXTD é uma variação do SAX que utiliza um símbolo para representar o valor médio e outro para representa a tendência de cada segmento.

O SAX é uma das técnicas de representações simbólicas amplamente utilizadas em mineração de séries temporais. No entanto, esta técnica não foi a mais selecionada para compor o ensemble. Isto deve-se ao fato que o batimento cardíaco com janela do QRS igual

Tabela 17 – Resultado obtido pelo Ensemble com MPGA usando a base de dados MIT

Resultados do Ensemble por otimização global					
Repr.	Tx. treino	Tx. teste	EER	Wilcoxon	Parâmetro
EFD	99.8536	99.4395	96.8610	V - 0.0168	$\lambda = 126; \mu = 4; \alpha = 60; \nu = 1; \kappa = 3$
EFVD	98.6616	97.5336	84.3049	V - 0.0001	$\lambda = 100; \mu = 3; \alpha = 46; \nu = 1; \kappa = 4$
EWD	99.8327	99.8318	93.1614	F - 0.6264	$\lambda = 121; \mu = 4; \alpha = 44; \nu = 3; \kappa = 2$
SAX	99.6236	99.5516	92.7691	F - 0.0690	$\lambda = 120; \mu = 4; \alpha = 44; \nu = 6; \kappa = 1$
SAXTD	94.4166	91.8722	80.4372	V - 0.0001	$\lambda = 92; \mu = 1; \alpha = 43; \nu = 10; \kappa = 8$
Ensemble	99.9373	99.8879	97.6457	-	Representações acima

Resultados do Ensemble por otimização individual					
Repr.	Tx. treino	Tx. teste	EER	Wilcoxon	Parâmetro
EFD	99.9791	99.8879	86.5471	F - 0.6540	$\lambda = 110; \mu = 4; \alpha = 52; \nu = 1; \kappa = 2$
EFVD	99.9791	99.7758	85.2018	F - 0.5936	$\lambda = 118; \mu = 4; \alpha = 63; \nu = 1; \kappa = 2$
EWD	99.9373	99.7758	84.3610	F - 0.8888	$\lambda = 93; \mu = 4; \alpha = 28; \nu = 1; \kappa = 2$
SAX	99.6445	99.6637	91.4798	F - 0.7441	$\lambda = 125; \mu = 4; \alpha = 64; \nu = 2; \kappa = 2$
SAXTD	99.6654	99.6637	92.0404	F - 0.7441	$\lambda = 126; \mu = 4; \alpha = 46; \nu = 3; \kappa = 1$
Ensemble	99.9791	99.8318	94.4507	-	Representações acima

Fonte: Henrique Passos, 2018

Tabela 18 – Resultado obtido pelo Ensemble com MPGA usando a base de dados PTB

Resultados do Ensemble por otimização global					
Repr.	Tx. treino	Tx. teste	EER	Wilcoxon	Parâmetro
EFD	99.6297	99.1857	59.7576	V - 0.0002	$\lambda = 113; \mu = 4; \alpha = 47; \nu = 3; \kappa = 2$
EFVD	94.5931	96.5485	63.4126	V - 0.0002	$\lambda = 53; \mu = 4; \alpha = 56; \nu = 2; \kappa = 2$
ESAX	99.4824	99.6854	66.2534	V - 0.0002	$\lambda = 128; \mu = 4; \alpha = 37; \nu = 8; \kappa = 2$
EWD	99.8678	99.7502	94.4665	V - 0.0097	$\lambda = 127; \mu = 4; \alpha = 29; \nu = 1; \kappa = 1$
SAXTD	99.0365	99.3338	75.8767	V - 0.0002	$\lambda = 128; \mu = 3; \alpha = 26; \nu = 8; \kappa = 2$
Ensemble	99.9584	99.9260	95.0495	-	Representações acima

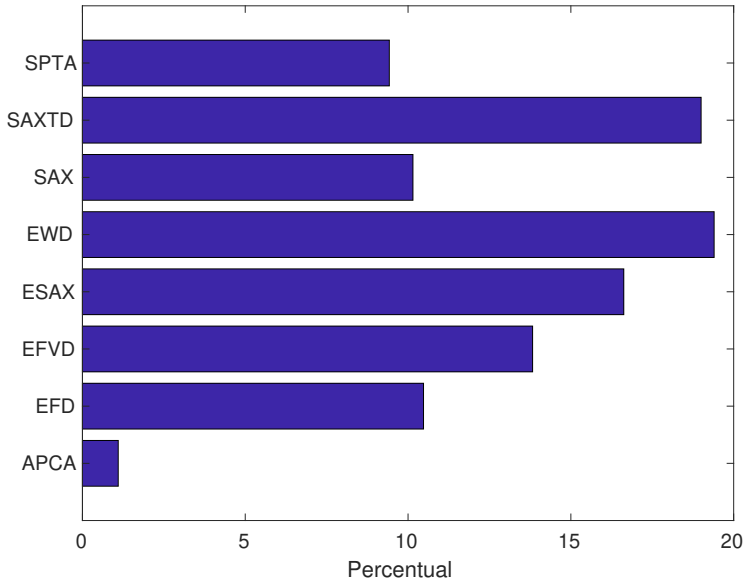
Resultados do Ensemble por otimização individual					
Repr.	Tx. treino	Tx. teste	EER	Wilcoxon	Parâmetro
EFD	99.6977	99.5281	60.5719	V - 0.0005	$\lambda = 128; \mu = 4; \alpha = 43; \nu = 5; \kappa = 2$
EFVD	99.7393	99.5651	39.8723	V - 0.0002	$\lambda = 123; \mu = 4; \alpha = 51; \nu = 4; \kappa = 1$
ESAX	99.5201	99.7039	84.1954	V - 0.0058	$\lambda = 118; \mu = 4; \alpha = 62; \nu = 1; \kappa = 1$
EWD	99.9244	99.8890	91.5703	F - 0.2343	$\lambda = 56; \mu = 4; \alpha = 55; \nu = 1; \kappa = 1$
SAXTD	99.5126	99.7224	79.7076	V - 0.0036	$\lambda = 127; \mu = 4; \alpha = 56; \nu = 1; \kappa = 2$
Ensemble	99.9471	99.9352	92.0607	-	Representações acima

Fonte: Henrique Passos, 2018

a 128 não é altamente gaussiano, conforme ilustrado na figura 32. Desta forma, o SAX perde um pouco sua capacidade de representação, uma vez que na etapa de simbolização assume que os dados são gaussianos para definição das áreas equiprováveis.

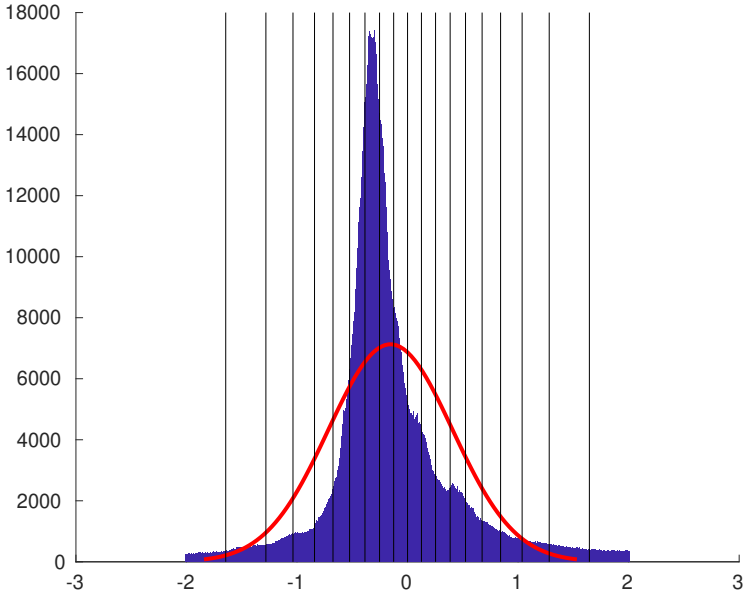
O EFD e EWD são técnicas de representação simbólica muito similar, porém a diferença reside na forma de simbolizar os valores. De forma a compreender melhor a diferença entre estas duas técnicas, considere o histograma dos batimentos cardíacos com janela do QRS igual a 128 e com linhas verticais representando a delimitação das regiões de simbolização de cada técnica, conforme ilustrado nas figuras 33 e 34. Neste caso, é nítido a diferença entre as duas simbolizações. No caso do EWD a simbolização está concentrada

Figura 31 – Frequência percentual com qua cada técnica foi selecionada pelos algoritmos evolutivos



Fonte: Henrique Passos, 2018

Figura 32 – Histograma dos batimentos cardíacos com janela do QRS igual a 128 e sua aproximação por uma gaussiana



Fonte: Henrique Passos, 2018

Tabela 19 – Resultado obtido pelo Ensemble com MPGA usando a base de dados MIT

Resultados do Ensemble por otimização global					
Repr.	Tx. treino	Tx. teste	EER	Wilcoxon	Parâmetro
EFVD	99.9582	99.7758	81.3341	F - 0.7978	$\lambda = 122; \mu = 4; \alpha = 56; \nu = 1; \kappa = 1$
ESAX	96.4659	95.6278	81.8946	V - 0.0001	$\lambda = 67; \mu = 2; \alpha = 10; \nu = 5; \kappa = 3$
EWD	99.8327	99.7758	88.7892	F - 0.6895	$\lambda = 106; \mu = 4; \alpha = 33; \nu = 2; \kappa = 1$
SAXTD	99.3308	98.5426	91.2556	V - 0.0023	$\lambda = 123; \mu = 3; \alpha = 28; \nu = 1; \kappa = 1$
SPTA	42.4927	32.0628	16.5359	V - 0.0001	$\lambda = 108; \mu = 4; \alpha = 59; \nu = 7; \kappa = 6$
Ensemble	99.9791	99.7197	94.8991	-	Representações acima

Resultados do Ensemble por otimização individual					
Repr.	Tx. treino	Tx. teste	EER	Wilcoxon	Parâmetro
EFVD	99.9791	99.5516	77.8587	F - 0.1204	$\lambda = 102; \mu = 4; \alpha = 46; \nu = 1; \kappa = 1$
ESAX	99.7072	99.6637	91.6480	F - 0.4562	$\lambda = 128; \mu = 4; \alpha = 31; \nu = 7; \kappa = 1$
EWD	99.9791	99.8318	92.0964	F - 0.5843	$\lambda = 119; \mu = 4; \alpha = 5; \nu = 1; \kappa = 1$
SAXTD	99.6654	99.6076	92.3206	F - 0.2083	$\lambda = 112; \mu = 4; \alpha = 31; \nu = 3; \kappa = 1$
SPTA	78.4191	72.5336	53.5314	V - 0.0001	$\lambda = 31; \mu = 4; \alpha = 9; \nu = 1; \kappa = 1$
Ensemble	99.9791	99.8879	95.4036	-	Representações acima

Fonte: Henrique Passos, 2018

Tabela 20 – Resultado obtido pelo Ensemble com MPGA usando a base de dados PTB

Resultados do Ensemble por otimização global					
Repr.	Tx. treino	Tx. teste	EER	Wilcoxon	Parâmetro
EFVD	99.6902	99.5373	50.9485	V - 0.0006	$\lambda = 99; \mu = 4; \alpha = 63; \nu = 4; \kappa = 1$
ESAX	99.4672	99.6669	77.5238	F - 0.0883	$\lambda = 123; \mu = 4; \alpha = 57; \nu = 3; \kappa = 1$
EWD	99.8715	99.7502	93.6800	F - 0.1846	$\lambda = 115; \mu = 4; \alpha = 64; \nu = 1; \kappa = 1$
SAXTD	98.9987	99.2320	81.2992	V - 0.0036	$\lambda = 98; \mu = 3; \alpha = 24; \nu = 5; \kappa = 2$
SPTA	44.0679	35.6621	22.8371	V - 0.0002	$\lambda = 117; \mu = 1; \alpha = 28; \nu = 2; \kappa = 18$
Ensemble	99.9547	99.8334	93.7911	-	Representações acima

Resultados do Ensemble por otimização individual					
Repr.	Tx. treino	Tx. teste	EER	Wilcoxon	Parâmetro
EFVD	99.6977	99.5558	59.9426	V - 0.0002	$\lambda = 126; \mu = 4; \alpha = 59; \nu = 5; \kappa = 2$
ESAX	99.5164	99.7039	77.8199	V - 0.0051	$\lambda = 125; \mu = 4; \alpha = 62; \nu = 3; \kappa = 1$
EWD	99.9282	99.8982	90.8208	F - 0.8479	$\lambda = 56; \mu = 4; \alpha = 52; \nu = 1; \kappa = 2$
SAXTD	99.5050	99.7409	89.0719	V - 0.0137	$\lambda = 128; \mu = 4; \alpha = 59; \nu = 1; \kappa = 1$
SPTA	84.0248	79.1894	56.2506	V - 0.0002	$\lambda = 59; \mu = 4; \alpha = 13; \nu = 1; \kappa = 1$
Ensemble	99.8904	99.8890	92.5789	-	Representações acima

Fonte: Henrique Passos, 2018

nas regiões com maior quantidade de dados, sendo este o fator de maior relevância para o desempenho dessa técnica. Por outro lado, a Discretização de Largura Igual obteve maior desempenho quando comparado com a Discretização de Frequência Igual, como evidenciado pela tabela 31.

6.5 Comparação de resultados

Comparando os resultados obtidos pela abordagem proposta com estudos publicados que utilizaram o conjunto de dados PTB, o desempenho obtido com ensemble de representa-

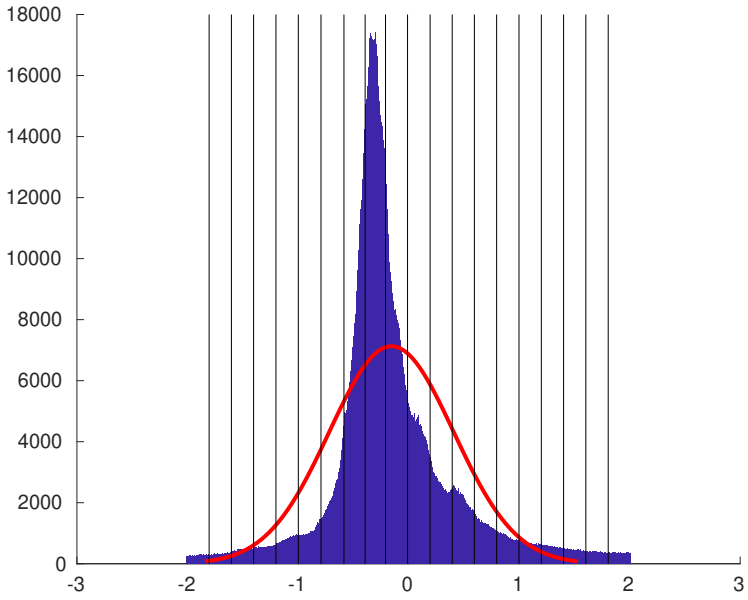
Tabela 21 – Resultado obtido pelo Ensemble com MPGA usando a base de dados MIT

Resultados do Ensemble por otimização global					
Repr.	Tx. treino	Tx. teste	EER	Wilcoxon	Parâmetro
EFD	99.7281	98.8229	91.6480	V - 0.0049	$\lambda = 126; \mu = 3; \alpha = 64; \nu = 4; \kappa = 1$
EFVD	99.9373	99.5516	77.0179	V - 0.0165	$\lambda = 99; \mu = 4; \alpha = 46; \nu = 1; \kappa = 2$
ESAX	99.6027	99.6637	91.7601	F - 0.1012	$\lambda = 116; \mu = 4; \alpha = 18; \nu = 5; \kappa = 2$
EWD	95.7131	92.9372	84.2489	V - 0.0001	$\lambda = 89; \mu = 4; \alpha = 37; \nu = 4; \kappa = 18$
SPTA	54.8306	44.9552	33.9126	V - 0.0001	$\lambda = 39; \mu = 2; \alpha = 38; \nu = 1; \kappa = 22$
Ensemble	99.9791	99.9439	94.8991	-	Representações acima

Resultados do Ensemble por otimização individual					
Repr.	Tx. treino	Tx. teste	EER	Wilcoxon	Parâmetro
EFD	99.9791	99.8879	85.4260	F - 0.5823	$\lambda = 110; \mu = 4; \alpha = 59; \nu = 1; \kappa = 2$
EFVD	99.9791	99.8318	87.7242	F - 1.0000	$\lambda = 101; \mu = 4; \alpha = 64; \nu = 1; \kappa = 1$
ESAX	99.7072	99.6637	91.6480	F - 0.7440	$\lambda = 128; \mu = 4; \alpha = 31; \nu = 7; \kappa = 1$
EWD	99.9373	99.7758	85.2018	F - 0.9256	$\lambda = 120; \mu = 4; \alpha = 48; \nu = 1; \kappa = 1$
SPTA	78.4191	72.5336	53.5314	V - 0.0001	$\lambda = 31; \mu = 4; \alpha = 9; \nu = 1; \kappa = 1$
Ensemble	99.9791	99.8318	94.0583	-	Representações acima

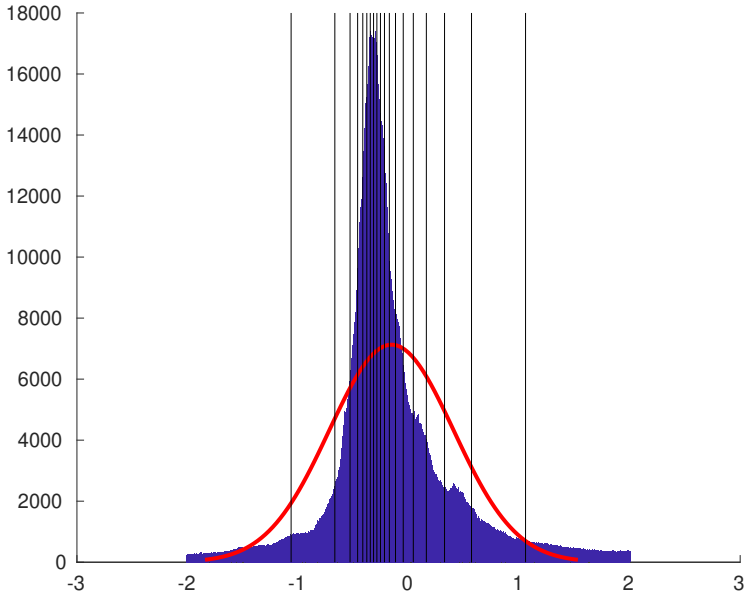
Fonte: Henrique Passos, 2018

Figura 33 – Histograma dos batimentos cardíacos com janela do QRS igual a 128 e sua aproximação por uma gaussiana, com linhas verticais representando as regiões de simbolização do EFD



Fonte: Henrique Passos, 2018

Figura 34 – Histograma dos batimentos cardíacos com janela do QRS igual a 128 e sua aproximação por uma gaussiana, com linhas verticais representando as regiões de simbolização do EWD



Fonte: Henrique Passos, 2018

ção simbólica é bastante satisfatório. Plataniotis, Hatzinakos e Lee (2006) alcançaram uma acurácia de 100% no reconhecimento de 14 indivíduos da base de dados PTB, utilizando características fiduciais. O ensemble de representação simbólica alcançou resultados muito próximo usando 290 indivíduos do PTB, o que corresponde a um conjunto de dados vinte vezes maior em relação ao de Plataniotis, Hatzinakos e Lee (2006). O tamanho do conjunto de dados está diretamente relacionado com a dificuldade na classificação, quanto maior a quantidade de indivíduos maior a dificuldade no reconhecimento dos batimentos cardíacos. (AGHAKABI; ZOKAEE, 2011) utilizou todos os indivíduos do conjunto de dados do PTB e alcançou uma acurácia igual a 94.7% usando o classificador KNN. Neste caso, a abordagem proposta nessa dissertação apresentou resultados superiores.

Muitas abordagens têm sido propostas para reconhecimento biométrico baseado em batimentos cardíacos. No entanto, não existe um consenso sobre a metodologia mais adequada para avaliar as abordagens. Parte dos trabalhos empregam conjunto de dados privado ou conjunto de dados públicos com a quantidade de indivíduos reduzida, dificultando um pouco a comparação. No entanto, ao comparar o desempenho da abordagem proposta com outros estudos é notável o desempenho no reconhecimento dos batimentos cardíacos e a sua capacidade de generalização. Um dos maiores problemas do emprego de representação simbólica é encontrar a parametrização mais adequada para um problema específico. A

adoção de algoritmos evolutivos tanto para a busca da melhor parametrização quando para a escolha das melhores técnicas a serem utilizadas no ensemble foi importante para obtenção dos resultados. As diversas execuções mostraram que existem várias configurações de ensemble com desempenho similar. Não houve nenhuma configuração que se destacasse, todas as configurações produziram bons resultados.

6.6 Considerações Finais

Essa dissertação propôs e avaliou ensemble de representação simbólica para reconhecimento biométrico baseado em sinais de ECG. As técnicas de representação simbólicas foram aplicadas para extração de características não fiduciais dos batimentos cardíacos. A base de dados PTB possuía 290 indivíduos, da qual foram extraídos aproximadamente 26400 batimentos cardíacos. A abordagem proposta obteve uma taxa de reconhecimento igual a 99.9352% sobre o conjunto de teste. Todas as configurações de ensemble alcançaram taxas de reconhecimento superiores a 99.7%.

Algumas técnicas de representação simbólica, tais como EWD e o SAXTD, quando sintonizadas adequadamente produziram resultados próximos aos obtidos pelo ensemble. O EWD se destacou pela sua forma de simbolização, visto que a diferença entre o EWD e EFD é apenas na forma de simbolização. EWD alcançou os melhores resultados individuais em comparação com as outras técnicas de representação simbólica. O SAXTD, o qual consiste de uma variação do SAX que codifica a tendência de cada segmento, alcançou melhores resultados em comparação do SAX. Isso mostra que a tendência codificada pelo SAXTD teve grande importância para o reconhecimento de batimentos cardíacos.

A busca de parâmetros para os componentes do ensemble pelos diversos algoritmos evolutivos produziu resultados similares. Isto mostra que a variação entre os algoritmos evolutivos não influenciou no processo de busca. Diversos resultados foram encontrados, mas olhando para os resultados finais de cada ensemble é notável a proximidade entre as parametrizações.

Analisando o parâmetro tamanho do alfabeto (α), é possível notar que grande parte dos valores obtidos para este parâmetro tendem a 64. Esse padrão pode ser observado em ambas as bases. Entretanto, as tabelas 7, 9, 13 e 19 apresentam exceções para esse comportamento. Isso mostra que variando apenas o valor de α de forma crescente, a taxa

de reconhecimento tem uma tendência crescente, porém podem existir picos de taxa de reconhecimento onde o α não é grande.

O ensemble de representação simbólica mostrou se bastante promissor para o reconhecimento baseado em batimentos cardíacos, visto que o desempenho alcançado foi superior a vários trabalhos publicados. Resultados individuais mostram que nem sempre uma única representação simbólica pode ser capaz altas taxas de reconhecimento, porém a composição de representação simbólica tem grande chance de produzir bons resultados. Isso deve-se ao fato que o ensemble contém diversas representações simbólicas que são mais indicados para resolver problemas distintos. Essa diversidade tem papel fundamental para o desempenho do ensemble.

7 Conclusão e Perspectivas Futuras

7.1 O enfoque da pesquisa

A quantidade de dados gerado com o passar dos anos tem aumentado drasticamente, isso demanda a necessidade de compactação e principalmente de técnicas de mineração de dados de forma a permitir uma melhor compreensão destes visando tomadas de decisões (MCAFEE et al., 2012). Existem dados no qual a ordem cronológica é fundamental para a sua compreensão, esse tipo de dados é chamado de séries temporais. Uma das áreas responsáveis pelo estudo desse tipo de dados é a mineração de dados em séries temporais (FU, 2011). Várias representações de séries temporais têm sido propostas em mineração de dados visando promover melhorias na redução da dimensionalidade, na simbolização dos dados e na resolução do problema. Uma das representações de maior destaque é o SAX, visto que diversas outras representações já foram derivadas desta técnica (KEOGH et al., 2001a; LIN et al., 2003; LKHAGVA; SUZUKI; KAWAGOE, 2006b). Por outro lado, sinais de ECG têm sido empregado com sucesso para reconhecimento biométrico Biel et al. (2001), Irvine et al. (2001), Kyoso e Uchiyama (2001). Alguns trabalhos têm investigado representações simbólicas para sinais de ECG visando monitoramento cardíaco Tayebi et al. (2011). No entanto, há poucos trabalhos que investigam o emprego de representações simbólicas para sinais de ECG visando reconhecimento biométrico.

Esse trabalho foca no emprego de representações simbólicas para sinais de ECG. O ensemble de técnicas de representação simbólica proposto nesse tem como objetivo combinar as sequencias simbólicas mais relevantes para a tarefa de reconhecimento biométrico. Para a criação do ensemble é necessário uma busca exploratória com o intuito de selecionar as melhores técnicas de representação simbólica. A busca das melhores técnicas de representação simbólicas e seus parâmetros foi realizada via algoritmos evolutivos. As representações simbólicas estudadas nesta dissertação foram 1DSAX, ASAX, DESAX, EFD, EFVD, ENSAX, ESAX, ESAX-Kmeans, EWD, GASAX, RSAX, SAX, SAX-Kmeans, SAXTD, SPTA e VWSAX.

7.2 Contribuições e resultados obtidos

Há diversas técnicas de representação simbólicas propostas na literatura, porém muitas ainda não tinham sido aplicadas no reconhecimento biométrico. Esse trabalho realizou uma análise detalhada do ensemble de representações simbólicas com o intuito de atacar o reconhecimento biométrico baseado em sinais de ECG. O ensemble de representações simbólicas proposto nesse trabalho é inspirado na ideia de comitê de máquinas de aprendizado. O ensemble foi proposto como uma forma de aliviar o processo de escolha da técnica de representação simbólica e de seus parâmetros. A seleção dos componentes e seus parâmetros foi realizada via algoritmos evolutivos. Os resultados alcançados demonstraram que a abordagem proposta é bastante promissora.

Três novas técnicas de representação simbólica foram propostas no decorrer desse trabalho, o ESAX-Kmeans, R-Kmeans e SAX-Kmeans. Um estudo comparativo envolvendo diversas técnicas de representação simbólica foi realizado e os resultados foram publicados. Nesse estudo comparativo, as técnicas propostas alcançaram resultados relevantes na tarefa de reconhecimento biométrico baseado em sinais de ECG. Além disso, esse trabalho apresentou no capitulo 3 uma descrição detalhada das principais técnicas de representação simbólicas.

Com base nos resultados obtidos, a abordagem proposta resultou em taxas de reconhecimento superiores a 99.7%, sendo que em alguns casos, o resultado alcançado foi superior aos obtidos individualmente pelos componentes do ensemble. Isso mostra a diversidade de representações que pode ser agregada para resolver um dado problema.

7.3 Perspectivas Futuras

Com relação a trabalhos futuros, pretende-se investigar o emprego de outros classificadores como Máquina de Vetores Suporte, Árvores de Decisão, Máquina de Aprendizado Extremo, etc. Outras medidas de distância podem ser adotadas no classificador KNN, tais como distância de Manhattan, distância de Mahalanobis, Duração do Tempo Dinâmico (do inglês Dynamic Time Warping - DTW), etc. Novos classificadores e outras medidas de distâncias podem resolver o problema de classificação de forma diferente e possivelmente melhorar o desempenho.

Modificações no ensemble de técnicas de representação simbólica podem ser adotadas, como por exemplo, pode ser inserido na codificação outros classificadores e seus parâmetros. A composição dos resultados do ensemble pode ser realizada por voto ponderado ou média ponderada. Outras técnicas de representação simbólica publicadas podem ser adicionadas ao ensemble. Além disso, representações não simbólicas também podem compor o ensemble.

Outro ponto interessante da abordagem proposta é que ao final da execução do algoritmo evolutivo informações importantes sobre as representações podem ser extraídas, por exemplo, as representações simbólicas mais frequentes no final da execução tendem a ser mais adequada para o problema tratado. Neste sentido, duas representações simbólicas, que estavam presente mais vezes nos cromossomos da última geração do algoritmo evolutivo, podem ser combinadas produzindo uma nova técnica de representação simbólica.

A abordagem proposta nesta dissertação pode ser aplicada a outros conjuntos de dados, tais como ECG-ID (LUGOVAYA, 2005) ou a outros sinais biológicos como o eletroencefalograma e o eletromiograma visando atacar uma determinada tarefa de classificação.

Referências[1]

ADDISON, P. S. Wavelet transforms and the ecg: a review. *Physiological measurement*, IOP Publishing, v. 26, n. 5, p. R155, 2005. Citado na página 27.

ADEOYE, O. S. A survey of emerging biometric technologies. *International Journal of Computer Applications*, International Journal of Computer Applications, 244 5 th Avenue,# 1526, New York, NY 10001, USA India, v. 10, 2010. Citado na página 17.

AGHAKABI, A.; ZOKAEE, S. Fusing dorsal hand vein and ecg for personal identification. In: IEEE. *Electrical and Control Engineering (ICECE), 2011 International Conference on*. [S.l.], 2011. p. 5933–5936. Citado na página 87.

AGRAFIOTI, F.; HATZINAKOS, D. Fusion of ecg sources for human identification. In: IEEE. *Communications, Control and Signal Processing, 2008. ISCCSP 2008. 3rd International Symposium on*. [S.l.], 2008. p. 1542–1547. Citado na página 65.

AHA, D. W.; KIBLER, D.; ALBERT, M. K. Instance-based learning algorithms. *Machine learning*, Springer, v. 6, n. 1, p. 37–66, 1991. Citado na página 65.

AL-RAISI, A. N.; AL-KHOURI, A. M. Iris recognition and the challenge of homeland and border control security in uae. *Telematics and Informatics*, Elsevier, v. 25, n. 2, p. 117–132, 2008. Citado na página 15.

ALLEN, P. *Calais migrants mutilate fingerprints to hide true identity, Daily Mail (n.d.)*. 2009. Disponível em: <http://www.dailymail.co.uk/news/worldnews/article-1201126/Calais-migrantsmutilate-fingertips-hide-true-identity.html>. Citado na página 17.

BAI, X. et al. Time series representation: a random shifting perspective. In: *Web-Age Information Management*. [S.l.]: Springer, 2013. p. 37–50. Citado na página 42.

BARNAGHI, P.; BAKAR, A. A.; OTHMAN, Z. A. Enhanced symbolic aggregate approximation method for financial time series data representation. In: IEEE. *Information Science and Service Science and Data Mining (ISSDM), 2012 6th International Conference on New Trends in*. [S.l.], 2012. p. 790–795. Citado na página 39.

BARNAGHI, P. M.; BAKAR, A. A.; OTHMAN, Z. A. Enhanced symbolic aggregate approximation (en-sax) as an improved representation method for financial time series data. *International Journal of Soft Computing*, v. 8, n. 4, p. 261–268, 2013. Citado na página 39.

BIEL, L. et al. Ecg analysis: a new approach in human identification. *Instrumentation and Measurement, IEEE Transactions on*, IEEE, v. 50, n. 3, p. 808–812, 2001. Citado 3 vezes nas páginas 18, 25 e 90.

BOUSSELJOT, R.; KREISELER, D.; SCHNABEL, A. Nutzung der ekg-signaldatenbank cardiodat der ptb über das internet. *Biomedizinische Technik/Biomedical Engineering*, v. 40, n. s1, p. 317–318, 1995. Citado na página 69.

CAI, Y.; NG, R. Indexing spatio-temporal trajectories with chebyshev polynomials. In: ACM. *Proceedings of the 2004 ACM SIGMOD international conference on Management of data*. [S.l.], 2004. p. 599–610. Citado na página 19.

[1] De acordo com a Associação Brasileira de Normas Técnicas. NBR 6023.

CAMERRA, A. et al. isax 2.0: Indexing and mining one billion time series. In: IEEE. *Data Mining (ICDM), 2010 IEEE 10th International Conference on*. [S.l.], 2010. p. 58–67. Citado na página 18.

CANENTO, F. et al. Review and comparison of real time electrocardiogram segmentation algorithms for biometric applications. In: *Proceedings of the 6th Int'l Conference on Health Informatics (HEALTHINF)*. [S.l.: s.n.], 2013. Citado na página 27.

CHAN, A. D. et al. Wavelet distance measure for person identification using electrocardiograms. *Instrumentation and Measurement, IEEE Transactions on*, IEEE, v. 57, n. 2, p. 248–253, 2008. Citado 2 vezes nas páginas 26 e 29.

CHAN, H.; BLEDSOE, W. A man-machine facial recognition system: some preliminary results. *Panoramic Research Inc*, 1965. Citado na página 16.

CHAN, K.-P.; FU, A. W.-C. Efficient time series matching by wavelets. In: IEEE. *Data Engineering, 1999. Proceedings., 15th International Conference on*. [S.l.], 1999. p. 126–133. Citado na página 19.

CHEN, Q. et al. Indexable pla for efficient similarity search. In: VLDB ENDOWMENT. *Proceedings of the 33rd international conference on Very large data bases*. [S.l.], 2007. p. 435–446. Citado na página 18.

CHIU, C.-C.; CHUANG, C.-M.; HSU, C.-Y. A novel personal identity verification approach using a discrete wavelet transform of the ecg signal. In: IEEE. *Multimedia and Ubiquitous Engineering, 2008. MUE 2008. International Conference on*. [S.l.], 2008. p. 201–206. Citado na página 28.

CHUN, S. Y. et al. Ecg based user authentication for wearable devices using short time fourier transform. In: IEEE. *Telecommunications and Signal Processing (TSP), 2016 39th International Conference on*. [S.l.], 2016. p. 656–659. Citado na página 26.

CLERC, M.; KENNEDY, J. The particle swarm-explosion, stability, and convergence in a multidimensional complex space. *IEEE transactions on Evolutionary Computation*, IEEE, v. 6, n. 1, p. 58–73, 2002. Citado 2 vezes nas páginas 53 e 56.

COELHO, A. L.; LIMA, C. A.; ZUBEN, F. V. Ga-based selection of components for heterogeneous ensembles of support vector machines. In: IEEE. *Evolutionary Computation, 2003. CEC'03. The 2003 Congress on*. [S.l.], 2003. v. 3, p. 2238–2245. Citado na página 61.

COVER, T.; HART, P. Nearest neighbor pattern classification. *IEEE transactions on information theory*, IEEE, v. 13, n. 1, p. 21–27, 1967. Citado na página 65.

DAN, J. et al. Piecewise trend approximation: a ratio-based time series representation. In: HINDAWI PUBLISHING CORPORATION. *Abstract and Applied Analysis*. [S.l.], 2013. v. 2013. Citado na página 44.

DAUGMAN, J. New methods in iris recognition. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, v. 37, n. 5, p. 1167–1175, 2007. Citado na página 15.

DAUGMAN, J. G. High confidence visual recognition of persons by a test of statistical independence. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, v. 15, n. 11, p. 1148–1161, 1993. Citado na página 16.

DING, H. et al. Querying and mining of time series data: experimental comparison of representations and distance measures. *Proceedings of the VLDB Endowment*, VLDB Endowment, v. 1, n. 2, p. 1542–1552, 2008. Citado na página 18.

DOUGHERTY, J. et al. Supervised and unsupervised discretization of continuous features. In: *Machine learning: proceedings of the twelfth international conference*. [S.l.: s.n.], 1995. v. 12, p. 194–202. Citado na página 36.

DUBIN, D. *Rapid Interpretation of EKG's*. [S.l.]: Cover Publishing Company, 1989. v. 200. Citado na página 24.

EBERHART, R.; KENNEDY, J. A new optimizer using particle swarm theory. In: IEEE. *Micro Machine and Human Science, 1995. MHS'95., Proceedings of the Sixth International Symposium on*. [S.l.], 1995. p. 39–43. Citado na página 53.

ESBENSEN, K.; SCHÖNKOPF, S.; MIDTGAARD, T. Multivariate anal. *Practice, 1st ed. Trondheim, Norway: Camo*, v. 1, p. 361, 1994. Citado na página 26.

ESLING, P.; AGON, C. Time-series data mining. *ACM Computing Surveys (CSUR)*, ACM, v. 45, n. 1, p. 12, 2012. Citado na página 18.

FALOUTSOS, C.; RANGANATHAN, M.; MANOLOPOULOS, Y. *Fast subsequence matching in time-series databases*. [S.l.]: ACM, 1994. v. 23. Citado na página 19.

FATEMIAN, S. Z.; AGRAFIOTI, F.; HATZINAKOS, D. Heartid: Cardiac biometric recognition. In: IEEE. *Biometrics: Theory Applications and Systems (BTAS), 2010 Fourth IEEE International Conference on*. [S.l.], 2010. p. 1–5. Citado na página 18.

FERN, X. Z.; BRODLEY, C. E. Solving cluster ensemble problems by bipartite graph partitioning. In: ACM. *Proceedings of the twenty-first international conference on Machine learning*. [S.l.], 2004. p. 36. Citado na página 61.

FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. *The elements of statistical learning*. [S.l.]: Springer series in statistics New York, 2001. v. 1. Citado na página 60.

FU, T.-c. A review on time series data mining. *Engineering Applications of Artificial Intelligence*, Elsevier, v. 24, n. 1, p. 164–181, 2011. Citado 2 vezes nas páginas 18 e 90.

FU, T.-c. et al. Representing financial time series based on data point importance. *Engineering Applications of Artificial Intelligence*, Elsevier, v. 21, n. 2, p. 277–300, 2008. Citado na página 19.

FUAD, M. M. M. Differential evolution versus genetic algorithms: Towards symbolic aggregate approximation of non-normalized time series. In: ACM. *Proceedings of the 16th International Database Engineering & Applications Sysmposium*. [S.l.], 2012. p. 205–210. Citado 2 vezes nas páginas 38 e 39.

FUAD, M. M. M. *Genetic algorithms-based symbolic aggregate approximation*. [S.l.]: Springer, 2012. Citado na página 38.

GALBALLY-HERRERO, J. et al. On the vulnerability of fingerprint verification systems to fake fingerprints attacks. In: IEEE. *Carnahan Conferences Security Technology, Proceedings 2006 40th Annual IEEE International*. [S.l.], 2006. p. 130–136. Citado na página 17.

GEURTS, P. Pattern extraction for time series classification. In: SPRINGER. *European Conference on Principles of Data Mining and Knowledge Discovery*. [S.l.], 2001. p. 115–127. Citado na página 18.

GOLDBERGER, A. L. et al. Physiobank, physiotoolkit, and physionet components of a new research resource for complex physiologic signals. *Circulation*, Am Heart Assoc, v. 101, n. 23, p. e215–e220, 2000. Citado 2 vezes nas páginas 69 e 70.

GOMES, C. M. G. F. *Autenticação Biométrica por Reconhecimento de Voz*. Tese (Doutorado) — Universidade do Minho, Portugal, 2007. Citado na página 15.

GOULD, S. J. Evolution's erratic pace. *Natural History*, ERIC, v. 86, n. 5, p. 12–16, 1977. Citado na página 51.

GÜLER, İ.; ÜBEYLI, E. D. A modified mixture of experts network structure for ecg beats classification with diverse features. *Engineering Applications of Artificial Intelligence*, Elsevier, v. 18, n. 7, p. 845–856, 2005. Citado na página 60.

GULLO, F. et al. A time series representation model for accurate and fast similarity detection. *Pattern Recognition*, Elsevier, v. 42, n. 11, p. 2998–3014, 2009. Citado na página 19.

HADJITODOROV, S. T.; KUNCHEVA, L. I.; TODOROVA, L. P. Moderate diversity for better cluster ensembles. *Information Fusion*, Elsevier, v. 7, n. 3, p. 264–275, 2006. Citado na página 61.

HAMILTON, P. S.; TOMPKINS, W. J. Quantitative investigation of qrs detection rules using the mit/bih arrhythmia database. *IEEE transactions on biomedical engineering*, IEEE, n. 12, p. 1157–1165, 1986. Citado na página 17.

HASHEM, S. Optimal linear combinations of neural networks. *Neural networks*, Elsevier, v. 10, n. 4, p. 599–614, 1997. Citado 2 vezes nas páginas 60 e 62.

HATWAR, K. S.; BADHIYE, S. S. Alphabetic time series representation using trend based approach. In: IEEE. *Innovations in Information, Embedded and Communication Systems (ICIIECS), 2015 International Conference on*. [S.l.], 2015. p. 1–4. Citado na página 44.

HAYKIN, S. S. et al. *Neural networks and learning machines*. [S.l.]: Pearson Education Upper Saddle River, 2009. v. 3. Citado 2 vezes nas páginas 18 e 20.

HOFFMANN, U. et al. An efficient p300-based brain–computer interface for disabled subjects. *Journal of Neuroscience methods*, Elsevier, v. 167, n. 1, p. 115–125, 2008. Citado na página 17.

HOLLAND, J. H. *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. [S.l.]: MIT press, 1992. Citado na página 49.

HOLLANDER, M.; WOLFE, D. A.; CHICKEN, E. *Nonparametric statistical methods*. [S.l.]: John Wiley & Sons, 2013. Citado na página 71.

HYNDMAN, R. J.; ATHANASOPOULOS, G. *Forecasting: principles and practice*. [S.l.]: OTexts, 2014. Citado na página 36.

IRVINE, J. et al. Heart rate variability: a new biometric for human identification. In: *Proceedings of the International Conference on Artificial Intelligence (IC-AI'01)*. [S.l.: s.n.], 2001. p. 1106–1111. Citado 2 vezes nas páginas 18 e 90.

IRVINE, J. M.; ISRAEL, S. A. A sequential procedure for individual identity verification using ecg. *EURASIP Journal on Advances in Signal Processing*, Springer, v. 2009, n. 1, p. 243215, 2009. Citado na página 66.

IRVINE, J. M. et al. eigenpulse: Robust human identification from cardiovascular function. *Pattern Recognition*, Elsevier, v. 41, n. 11, p. 3427–3435, 2008. Citado na página 66.

ISRAEL, S. A. et al. Ecg to identify individuals. *Pattern recognition*, Elsevier, v. 38, n. 1, p. 133–142, 2005. Citado 2 vezes nas páginas 17 e 66.

ISRAEL, S. A. et al. Fusing face and ecg for personal identification. In: IEEE. *Applied Imagery Pattern Recognition Workshop, 2003. Proceedings. 32nd*. [S.l.], 2003. p. 226–231. Citado na página 18.

JACOBS, R. A. et al. Adaptive mixtures of local experts. *Neural computation*, MIT Press, v. 3, n. 1, p. 79–87, 1991. Citado na página 60.

JAIN, A. K.; FLYNN, P.; ROSS, A. A. *Handbook of biometrics*. [S.l.]: Springer Science & Business Media, 2007. Citado na página 15.

JAIN, A. K.; NANDAKUMAR, K.; ROSS, A. 50 years of biometric research: Accomplishments, challenges, and opportunities. *Pattern Recognition Letters*, Elsevier, v. 79, p. 80–105, 2016. Citado na página 16.

JAIN, A. K.; ROSS, A.; PRABHAKAR, S. An introduction to biometric recognition. *Circuits and Systems for Video Technology, IEEE Transactions on*, IEEE, v. 14, n. 1, p. 4–20, 2004. Citado 3 vezes nas páginas 22, 23 e 26.

KENNEDY, J. Small worlds and mega-minds: effects of neighborhood topology on particle swarm performance. In: IEEE. *Evolutionary Computation, 1999. CEC 99. Proceedings of the 1999 Congress on*. [S.l.], 1999. v. 3, p. 1931–1938. Citado na página 54.

KENNEDY, J.; EBERHART, R. C. A discrete binary version of the particle swarm algorithm. In: IEEE. *Systems, Man, and Cybernetics, 1997. Computational Cybernetics and Simulation., 1997 IEEE International Conference on*. [S.l.], 1997. v. 5, p. 4104–4108. Citado 2 vezes nas páginas 53 e 56.

KEOGH, E. et al. Dimensionality reduction for fast similarity search in large time series databases. *Knowledge and information Systems*, Springer, v. 3, n. 3, p. 263–286, 2001. Citado 4 vezes nas páginas 18, 31, 65 e 90.

KEOGH, E. et al. Locally adaptive dimensionality reduction for indexing large time series databases. *ACM SIGMOD Record*, ACM, v. 30, n. 2, p. 151–162, 2001. Citado 4 vezes nas páginas 18, 32, 44 e 45.

KEOGH, E.; KASETTY, S. On the need for time series data mining benchmarks: a survey and empirical demonstration. *Data Mining and knowledge discovery*, Springer, v. 7, n. 4, p. 349–371, 2003. Citado 2 vezes nas páginas 31 e 32.

KORN, F.; JAGADISH, H. V.; FALOUTSOS, C. Efficiently supporting ad hoc queries in large datasets of time sequences. *ACM SIGMOD Record*, ACM, v. 26, n. 2, p. 289–300, 1997. Citado na página 19.

KOTSIANTIS, S.; KANELLOPOULOS, D. Discretization techniques: A recent survey. *GESTS International Transactions on Computer Science and Engineering*, v. 32, n. 1, p. 47–58, 2006. Citado na página 36.

KREISELER, D.; BOUSSELIOT, R. Automatisierte ekg-auswertung mit hilfe der ekg-signaldatenbank cardiodat der ptb. *Biomedizinische Technik/Biomedical Engineering*, v. 40, n. s1, p. 319–320, 1995. Citado na página 69.

KULAHCIOGLU, B.; OZDEMIR, S.; KUMOVA, B. Application of symbolic piecewise aggregate approximation (paa) analysis to ecg signals. In: CITESEER. *17th IASTED International Conference on Applied Simulation and Modelling*. [S.l.], 2008. Citado na página 18.

KUNCHEVA, L. I. *Combining pattern classifiers: methods and algorithms*. [S.l.]: John Wiley & Sons, 2004. Citado 2 vezes nas páginas 59 e 60.

KYOSO, M.; UCHIYAMA, A. Development of an ecg identification system. In: IEEE. *Engineering in Medicine and Biology Society, 2001. Proceedings of the 23rd Annual International Conference of the IEEE*. [S.l.], 2001. v. 4, p. 3721–3723. Citado 2 vezes nas páginas 18 e 90.

LANGDON, W. B.; POLI, R. *Foundations of genetic programming*. [S.l.]: Springer Science & Business Media, 2013. Citado na página 49.

LAXMAN, S.; SASTRY, P. S. A survey of temporal data mining. *Sadhana*, Springer, v. 31, n. 2, p. 173–198, 2006. Citado na página 31.

LI, G.; ZHANG, L.; YANG, L. Tsx: A novel symbolic representation for financial time series. In: *PRICAI 2012: Trends in Artificial Intelligence*. [S.l.]: Springer, 2012. p. 262–273. Citado na página 37.

LIMA, C. A.; COELHO, A. L.; ZUBEN, F. J. V. Hybridizing mixtures of experts with support vector machines: Investigation into nonlinear dynamic systems identification. *Information Sciences*, Elsevier, v. 177, n. 10, p. 2049–2074, 2007. Citado na página 60.

LIMA, C. A. M.; COELHO, A.; ZUBEN, F. J. V. Ensembles of support vector machines for regression problems. In: IEEE. *Neural Networks, 2002. IJCNN'02. Proceedings of the 2002 International Joint Conference on*. [S.l.], 2002. v. 3, p. 2381–2386. Citado na página 61.

LIN, J. et al. A symbolic representation of time series, with implications for streaming algorithms. In: ACM. *Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*. [S.l.], 2003. p. 2–11. Citado 2 vezes nas páginas 33 e 90.

LIN, J. et al. Experiencing sax: a novel symbolic representation of time series. *Data Mining and knowledge discovery*, Springer, v. 15, n. 2, p. 107–144, 2007. Citado 3 vezes nas páginas 18, 19 e 32.

LKHAGVA, B.; SUZUKI, Y.; KAWAGOE, K. Extended sax: Extension of symbolic aggregate approximation for financial time series data representation. *DEWS2006 4A-i8*, v. 7, 2006. Citado na página 34.

LKHAGVA, B.; SUZUKI, Y.; KAWAGOE, K. New time series data representation esax for financial applications. In: *22nd International Conference on Data Engineering Workshops (ICDEW'06)*. [S.l.: s.n.], 2006. Citado 5 vezes nas páginas 18, 34, 35, 47 e 90.

LOURENÇO, A.; SILVA, H.; FRED, A. Unveiling the biometric potential of finger-based ecg signals. *Computational intelligence and neuroscience*, Hindawi Publishing Corp., v. 2011, p. 5, 2011. Citado na página 26.

LOURENÇO, A.; SILVA, H.; FRED, A. Ecg-based biometrics: A real time classification approach. In: IEEE. *Machine Learning for Signal Processing (MLSP), 2012 IEEE International Workshop on*. [S.l.], 2012. p. 1–6. Citado na página 27.

LUGOVAYA, T. Biometric human identification based on electrocardiogram. *LETI*, Saint-Petersburg, 2005. Citado 2 vezes nas páginas 28 e 92.

LUZ, E. J. d. S.; MENOTTI, D.; SCHWARTZ, W. R. Evaluating the use of ecg signal in low frequencies as a biometry. *Expert Systems with Applications*, Elsevier, v. 41, n. 5, p. 2309–2315, 2014. Citado na página 23.

MALTONI, D. et al. *Handbook of fingerprint recognition*. [S.l.]: Springer Science & Business Media, 2009. Citado na página 22.

MARCIALIS, G. L.; ROLI, F.; TIDU, A. Analysis of fingerprint pores for vitality detection. In: IEEE. *Pattern Recognition (ICPR), 2010 20th International Conference on*. [S.l.], 2010. p. 1289–1292. Citado na página 17.

MAUCERI, A. *FEASIBILITY STUDY OF PERSONNEL IDENTIFICATION BY SIGNATURE VERIFICATION*. [S.l.], 1965. Citado na página 16.

MCAFEE, A. et al. Big data: the management revolution. *Harvard business review*, v. 90, n. 10, p. 60–68, 2012. Citado na página 90.

MEDINA, A. J. R.; PULIDO, G. T.; RAMÍREZ-TORRES, J. G. A comparative study of neighborhood topologies for particle swarm optimizers. In: *IJCCI*. [S.l.: s.n.], 2009. p. 152–159. Citado na página 54.

MITCHELL, M. *An introduction to genetic algorithms*. [S.l.]: MIT press, 1998. Citado 2 vezes nas páginas 49 e 50.

MÜHLENBEIN, H.; SCHOMISCH, M.; BORN, J. The parallel genetic algorithm as function optimizer. *Parallel computing*, Elsevier, v. 17, n. 6-7, p. 619–632, 1991. Citado na página 51.

NILSSON, N. J. *Learning machines: foundations of trainable pattern-classifying systems*. [S.l.]: McGraw-Hill, 1965. Citado na página 19.

ODINAKA, I. et al. Ecg biometric recognition: A comparative analysis. *Information Forensics and Security, IEEE Transactions on*, IEEE, v. 7, n. 6, p. 1812–1824, 2012. Citado 3 vezes nas páginas 18, 27 e 65.

O'GORMAN, L. Comparing passwords, tokens, and biometrics for user authentication. *Proceedings of the IEEE*, IEEE, v. 91, n. 12, p. 2021–2040, 2003. Citado na página 15.

OTHERS and. *Hand ID system*. [S.l.]: Google Patents, 1971. US Patent 3,576,537. Citado na página 16.

PALANIAPPAN, R.; MANDIC, D. P. Biometrics from brain electrical activity: A machine learning approach. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, IEEE, v. 29, n. 4, p. 738–742, 2007. Citado na página 17.

PAN, J.; TOMPKINS, W. J. A real-time qrs detection algorithm. *Biomedical Engineering, IEEE Transactions on*, IEEE, n. 3, p. 230–236, 1985. Citado 2 vezes nas páginas 28 e 29.

PANKANTI, S.; BOLLE, R. M.; JAIN, A. Biometrics: The future of identification [guest eeditors' introduction]. *Computer*, IEEE, v. 33, n. 2, p. 46–49, 2000. Citado na página 17.

PASSOS, H. dos S. et al. Symbolic representations of time series applied to biometric recognition based on ecg signals. In: IEEE. *Neural Networks (IJCNN), 2017 International Joint Conference on*. [S.l.], 2017. p. 3199–3207. Citado 2 vezes nas páginas 34 e 45.

PERRONE, M. P.; COOPER, L. N. When networks disagree: Ensemble methods for hybrid neural networks. In: *How We Learn; How We Remember: Toward An Understanding Of Brain And Neural Systems: Selected Papers of Leon N Cooper*. [S.l.]: World Scientific, 1995. p. 342–358. Citado na página 64.

PHAM, N. D.; LE, Q. L.; DANG, T. K. Two novel adaptive symbolic representations for similarity search in time series databases. In: IEEE. *Web Conference (APWEB), 2010 12th International Asia-Pacific*. [S.l.], 2010. p. 181–187. Citado na página 35.

PHILLIPS, P. J. et al. The feret evaluation methodology for face-recognition algorithms. *IEEE Transactions on pattern analysis and machine intelligence*, IEEE, v. 22, n. 10, p. 1090–1104, 2000. Citado na página 15.

PLATANIOTIS, K. N.; HATZINAKOS, D.; LEE, J. K. Ecg biometric recognition without fiducial detection. In: IEEE. *Biometric Consortium Conference, 2006 Biometrics Symposium: Special Session on Research at the*. [S.l.], 2006. p. 1–6. Citado 2 vezes nas páginas 23 e 87.

PRUZANSKY, S. Pattern-matching procedure for automatic talker recognition. *The Journal of the Acoustical Society of America*, Acoustical Society of America, v. 35, n. 3, p. 354–358, 1963. Citado na página 16.

REPORT, T. U. F. C. *http://www.ftc.gov/sentinel/reports/sentinel-annual-reports/sentinel-cy2012.pdf*. 2012. Citado na página 15.

REVETT, K.; DERAVI, F.; SIRLANTZIS, K. Biosignals for user authentication-towards cognitive biometrics? In: IEEE. *Emerging Security Technologies (EST), 2010 International Conference on*. [S.l.], 2010. p. 71–76. Citado na página 17.

ROSS, A.; JAIN, A. Information fusion in biometrics. *Pattern recognition letters*, Elsevier, v. 24, n. 13, p. 2115–2125, 2003. Citado 2 vezes nas páginas 16 e 72.

SAHOO, J. P.; BEHERA, S.; ARI, S. A novel technique for qrs complex detection in ecg signal based on hilbert transform and autocorrelation. 2011. Citado na página 24.

SALGADO, R. M. et al. A hybrid ensemble model applied to the short-term load forecasting problem. In: IEEE. *Neural Networks, 2006. IJCNN'06. International Joint Conference on*. [S.l.], 2006. p. 2627–2634. Citado na página 61.

SHARKEY, A. J. *Combining artificial neural nets: ensemble and modular multi-net systems*. [S.l.]: Springer Science & Business Media, 2012. Citado na página 60.

SHEN, T.-W.; TOMPKINS, W.; HU, Y. One-lead ecg for identity verification. In: IEEE. *Engineering in medicine and biology, 2002. 24th annual conference and the annual fall meeting of the biomedical engineering society embs/bmes conference, 2002. proceedings of the second joint*. [S.l.], 2002. v. 1, p. 62–63. Citado na página 66.

SHEN, T.-W. D.; TOMPKINS, W. J.; HU, Y. H. Implementation of a one-lead ecg human identification system on a normal population. *Journal of Engineering and Computer Innovations*, Academic Journals, v. 2, n. 1, p. 12–21, 2010. Citado 2 vezes nas páginas 26 e 65.

SHI, Y.; EBERHART, R. A modified particle swarm optimizer. In: IEEE. *Evolutionary Computation Proceedings, 1998. IEEE World Congress on Computational Intelligence., The 1998 IEEE International Conference on*. [S.l.], 1998. p. 69–73. Citado na página 56.

SHIEH, J.; KEOGH, E. isax: disk-aware mining and indexing of massive time series datasets. *Data Mining and Knowledge Discovery*, Springer, v. 19, n. 1, p. 24–57, 2009. Citado na página 18.

SIMON, B. P.; ESWARAN, C. An ecg classifier designed using modified decision based neural networks. *Computers and Biomedical Research*, Elsevier, v. 30, n. 4, p. 257–272, 1997. Citado 3 vezes nas páginas 18, 23 e 25.

SIMON, M. et al. 1d-sax; a novel symbolic representation for time series. *IDA'13*, 2013. Citado na página 40.

SINGH, Y. N.; GUPTA, P. Ecg to individual identification. In: IEEE. *Biometrics: Theory, Applications and Systems, 2008. BTAS 2008. 2nd IEEE International Conference on*. [S.l.], 2008. p. 1–8. Citado na página 66.

SINGH, Y. N.; GUPTA, P. Biometrics method for human identification using electrocardiogram. In: SPRINGER. *International Conference on Biometrics*. [S.l.], 2009. p. 1270–1279. Citado na página 66.

SINGH, Y. N.; GUPTA, P. Correlation-based classification of heartbeats for individual identification. *Soft Computing*, Springer, v. 15, n. 3, p. 449–460, 2011. Citado 2 vezes nas páginas 17 e 66.

STARKWEATHER, T.; WHITLEY, D.; MATHIAS, K. Optimization using distributed genetic algorithms. In: SPRINGER. *International Conference on Parallel Problem Solving from Nature*. [S.l.], 1990. p. 176–185. Citado na página 51.

STORN, R. On the usage of differential evolution for function optimization. In: IEEE. *Fuzzy Information Processing Society, 1996. NAFIPS., 1996 Biennial Conference of the North American*. [S.l.], 1996. p. 519–523. Citado na página 52.

STORN, R.; PRICE, K. Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. *Journal of global optimization*, Springer, v. 11, n. 4, p. 341–359, 1997. Citado na página 52.

STREHL, A.; GHOSH, J. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research*, v. 3, n. Dec, p. 583–617, 2002. Citado na página 61.

SUBASI, A. Eeg signal classification using wavelet feature extraction and a mixture of expert model. *Expert Systems with Applications*, Elsevier, v. 32, n. 4, p. 1084–1093, 2007. Citado na página 60.

SUN, C. et al. Variance-wise segmentation for a temporal-adaptive sax. In: AUSTRALIAN COMPUTER SOCIETY, INC. *Proceedings of the Tenth Australasian Data Mining Conference-Volume 134*. [S.l.], 2012. p. 71–77. Citado na página 39.

SUN, S. Multitask learning for eeg-based biometrics. In: IEEE. *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*. [S.l.], 2008. p. 1–4. Citado na página 17.

SUN, Y. et al. An improvement of symbolic aggregate approximation distance measure for time series. *Neurocomputing*, Elsevier, v. 138, p. 189–198, 2014. Citado na página 42.

TAYEBI, H. et al. Ra-sax: resource-aware symbolic aggregate approximation for mobile ecg analysis. In: IEEE. *Mobile Data Management (MDM), 2011 12th IEEE International Conference on*. [S.l.], 2011. v. 1, p. 289–290. Citado na página 90.

THAKOR, N. V.; ZHU, Y.-S. Applications of adaptive filtering to ecg analysis: noise cancellation and arrhythmia detection. *IEEE transactions on biomedical engineering*, IEEE, v. 38, n. 8, p. 785–794, 1991. Citado na página 17.

TOLEDO, C. F. M. et al. Problema conjunto de dimensionamento de lotes e programação da produção. [sn], 2005. Citado na página 51.

TOPCHY, A.; JAIN, A. K.; PUNCH, W. Clustering ensembles: Models of consensus and weak partitions. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, v. 27, n. 12, p. 1866–1881, 2005. Citado na página 61.

TRAURING, M. Automatic comparison of finger-ridge patterns. *Nature*, v. 197, p. 938–940, 1963. Citado na página 16.

TRESP, V. *Committee machines*. [S.l.]: CRC Press Boca Raton, FL, 2001. 1–18 p. Citado 2 vezes nas páginas 59 e 61.

WANG, H. et al. An integrated biometric-based security framework using wavelet-domain hmm in wireless body area networks (wban). In: IEEE. *Communications (ICC), 2011 IEEE International Conference on*. [S.l.], 2011. p. 1–5. Citado na página 26.

WANG, Y. et al. Analysis of human electrocardiogram for biometric recognition. *EURASIP journal on Advances in Signal Processing*, Springer, v. 2008, n. 1, p. 148658, 2007. Citado na página 66.

WANG, Y.; TAN, T.; JAIN, A. K. Combining face and iris biometrics for identity verification. In: SPRINGER. *International Conference on Audio-and Video-Based Biometric Person Authentication*. [S.l.], 2003. p. 805–813. Citado na página 72.

WAYMAN, J. et al. An introduction to biometric authentication systems. *Biometric Systems*, Springer, p. 1–20, 2005. Citado 2 vezes nas páginas 15 e 16.

WU, Z.-H. Z. J.-X.; CHEN, Y. J. S.-F. Genetic algorithm based selective neural network ensemble. In: *IJCAI-01: proceedings of the Seventeenth International Joint Conference on Artificial Intelligence, Seattle, Washington*. [S.l.: s.n.], 2001. Citado na página 64.

WÜBBELER, G. et al. Verification of humans using the electrocardiogram. *Pattern Recognition Letters*, Elsevier, v. 28, n. 10, p. 1172–1175, 2007. Citado na página 28.

XI, X. et al. Fast time series classification using numerosity reduction. In: ACM. *Proceedings of the 23rd international conference on Machine learning*. [S.l.], 2006. p. 1033–1040. Citado na página 18.

XING, Z.; PEI, J.; KEOGH, E. A brief survey on sequence classification. *ACM Sigkdd Explorations Newsletter*, ACM, v. 12, n. 1, p. 40–48, 2010. Citado na página 65.

YEH, Y.-C.; WANG, W.-J. Qrs complexes detection for ecg signal: The difference operation method. *Computer methods and programs in biomedicine*, Elsevier, v. 91, n. 3, p. 245–254, 2008. Citado na página 28.

YI, B.-K.; FALOUTSOS, C. Fast time sequence indexing for arbitrary lp norms. In: VLDB. [S.l.], 2000. Citado na página 32.

ZALEWSKI, W. et al. Time series discretization based on the approximation of the local slope information. In: *Advances in Artificial Intelligence–IBERAMIA 2012*. [S.l.]: Springer, 2012. p. 91–100. Citado na página 36.

ZALEWSKI, W. et al. Symbolic representation based on temporal order information for time series classification. In: IEEE. *Intelligent Systems (BRACIS), 2013 Brazilian Conference on*. [S.l.], 2013. p. 95–100. Citado na página 36.

ZALEWSKI, W. et al. A symbolic representation method to preserve the characteristic slope of time series. In: SPRINGER. *SBIA*. [S.l.], 2012. p. 132–141. Citado na página 36.

ZHOU, Y.; KUMAR, A. Human identification using palm-vein images. *IEEE transactions on information forensics and security*, IEEE, v. 6, n. 4, p. 1259–1274, 2011. Citado na página 17.

ZHOU, Z.-H.; WU, J.; TANG, W. Ensembling neural networks: many could be better than all. *Artificial intelligence*, Elsevier, v. 137, n. 1, p. 239–263, 2002. Citado na página 20.

ZHOU, Z.-H. et al. Selectively ensembling neural classifiers. In: IEEE. *Neural Networks, 2002. IJCNN'02. Proceedings of the 2002 International Joint Conference on*. [S.l.], 2002. v. 2, p. 1411–1415. Citado na página 64.

ZIAUDDIN, S.; DAILEY, M. N. A robust hybrid iris localization technique. In: IEEE. *Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, 2009. ECTI-CON 2009. 6th International Conference on*. [S.l.], 2009. v. 2, p. 1058–1061. Citado na página 15.