



UNIVERSIDADE FEDERAL DO AMAZONAS
FACULDADE DE TECNOLOGIA
ENGENHARIA DA COMPUTAÇÃO

CLASSIFICAÇÃO DE SÉRIES TEMPORAIS COM APRENDIZAGEM PROFUNDA

CLEMILTON VASCONCELOS PEREIRA

MANAUS-AM

2018

CLEMILTON VASCONCELOS PEREIRA

CLASSIFICAÇÃO DE SÉRIES TEMPORAIS COM APRENDIZAGEM PROFUNDA

Monografia apresentada à Coordenação do
Curso de Engenharia da Computação da
Universidade Federal do Amazonas, como
parte dos requisitos necessários à obtenção
do título de Engenheiro de Computação.

Orientador: RAFAEL GIUSTI

MANAUS-AM

2018

Ficha Catalográfica

Ficha catalográfica elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

C957s Cruz Junior, António César Vieira da
Sistema de detecção de frequência fundamental para música /
António César Vieira da Cruz Junior. 2018
66 f.: il. color; 31 cm.

Orientador: Waldir Sabino da Silva Júnior
TCC de Graduação (Engenharia da Computação) - Universidade
Federal do Amazonas.

1. Frequência fundamental. 2. Processamento digital de áudio. 3.
Transcrição musical. 4. Detecção automática. I. Silva Júnior, Waldir
Sabino da II. Universidade Federal do Amazonas III. Título

CLASSIFICAÇÃO DE SÉRIES TEMPORAIS COM APRENDIZAGEM PROFUNDA

Clemilton Vasconcelos Pereira

MONOGRAFIA SUBMETIDA AO CORPO DOCENTE DO CURSO DE ENGENHARIA DA COMPUTAÇÃO DA UNIVERSIDADE FEDERAL DO AMAZONAS COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE ENGENHEIRO.

Aprovada por:

Prof. Rafael Giusti, D. Sc.

Prof., M. Sc.

Prof. , D. Sc.

Manaus
Julho de 2018

*Dedico este trabalho à minha mãe
Lúcia, "In memoriam", também aos meus
pais, Antonio e Elenir, e às minhas avós
Jacyra e Maria, meus exemplos de vida.*

Agradecimentos

Antes de tudo, a Deus, autor da vida, cujo a bondade e a misericórdia me seguem a cada dia. Tu és fiel, Senhor!

Aos meus pais, Antonio e Elenir, que sempre me apoiaram de todas as maneiras possíveis, à minha amada esposa Yasmin, sempre ao meu lado, e aos meus familiares, que sempre me incentivaram a persistir.

Ao meu orientador, prof. Dr. Waldir Sabino, pela paciência, disposição e direcionamento durante a execução deste projeto, e também pelos conselhos que levo para minha vida profissional.

A todos os professores da UFAM, em especial o prof. Dr. César Melo, aos amigos de turma Rosmael, Helton, João Victor, Gabriel, Jordan, Kabashima, Eduardo, Luiz Henrique, Micael e Clemilton, e aos demais amigos e colegas da UFAM pelo convívio, aprendizado e cooperação durante todos esses anos. Vocês foram essenciais nessa caminhada.

À família Machado, por todo o cuidado no momento em que mais precisei, e aos irmãos da Segunda Igreja Batista de Manaus, que me adotaram assim que cheguei nesta cidade. Também à Keyseane, que me ajudou em momentos importantes na escrita deste trabalho, e aos meus amigos José Eduardo, Vanessa e Débora.

A todos que, de algum modo, colaboraram para que este momento se concretizasse:

Muito obrigado!

“Se algum de vocês tem falta de sabedoria, peça-a a Deus, que a todos dá livremente, de boa vontade; e lhe será concedida.”

(Tiago 1.5)

Resumo

Existem diferentes caminhos para o reconhecimento automático de uma notas musical, e um deles é por meio da frequência fundamental (f_0). Muitas soluções e métodos de detecção automática de f_0 já foram desenvolvidas e apresentadas, obtendo resultados positivos. Todavia, é difícil que uma única solução de detecção seja de fato eficaz em ambientes muito diferentes daqueles para o qual a solução foi proposta. Diante disso, o presente projeto aborda o desenvolvimento de um sistema de detecção de f_0 para áudios musicais monofônicos, através de processamento digital de sinais, com o intuito de mapear os áudios, fornecendo as frequências fundamentais soadas em cada instante de tempo. Este trabalho também avaliou o sistema desenvolvido, por meio de experimentação a partir de uma base de dados construída para este fim. Percebeu-se que o sistema proposto apresenta boas respostas, sendo necessário melhorias em relação ao tratamento do efeito de ressonância.

Palavras-chave: Detecção automática de frequência fundamental, Processamento digital de áudio, Transcrição musical.

Abstract

There are different ways for a automatic recognition of musical note, one of them is through fundamental frequency (f_0). Many solutions and methods of automatic detection of f_0 already developed and presented, getting positive results. However, it's hard that one only solution has been effective in environments much different from these that it was proposed. Therefore, the present project approaches the development of a f_0 detector system for monophonic musical audios through digital signal processing, with the intention of mapping it, recording the fundamental frequencies sounded at every instant time. This work also evaluated the system developed, through experiments using a database built for it. It was perceived that the proposed system presents good answers, being necessary improvements in the treatment of the resonance effect.

Keywords: Automatic Detection of fundamental frequency, Digital Audio Processing, Musical transcription.

Lista de Figuras

2.1	Codificação One-hot	5
2.2	Princípio dos k-vizinhos mais próximos	6
2.3	Representação simplificada de um neurônio	7
2.4	Modelo matemático de um neurônio	8
2.5	Multilayer Perceptron. Cada círculo representa um neurônio mostrado anteriormente	9
2.6	Softmax vs ReLU. Imagem tirada do site <i>Towards Data Science</i> [1]	12
2.7	Imagem ilustrando o efeito do Dropout em uma rede neural onde os neurônios marcados foram desativados, forçando a rede não depender deles, evitando o <i>overfitting</i> [2].	14
2.8	Campo receptivo de um neurônio. Fonte: [3]	15
2.9	Disposição dos neurônios na camada de convolução e seu campo receptivo. Fonte: [4]	16
2.10	Neurônios espaçados por stride 1 e 2 das esquerda para a direita respectivamente. Separados à direita estão os pesos compartilhados entre todos os neurônios. Fonte: [5]	17
2.11	Representação de max-pooling e average-pooling. Fonte: [5]	18
2.12	Arquitetura LeNet-5. Fonte: [6]	19
2.13	Bloco Residual. Fonte: [7]	19
2.14	Distância Euclidiana. Fonte: [8]	22
2.15	A) Duas sequências Q e C são semelhantes, porém fora de fase. B) Para alinhar as sequências, construímos a matriz e procuramos pelo melhor caminho. C) O alinhamento resultado. Fonte: [9]	23
2.16	A similaridade entre as séries temporais A e B é reconhecida pela DTW. [8]	24

2.17	Visualização da janela de Sakoe-Chiba e do paralelogramo de Itakura.	
	Fonte: [10]	25
2.18	Visão geral de diferentes abordagens de aprendizado profundo para classi-	
	ficção de séries temporais. Fonte: [11]	26

Lista de Tabelas

2.1	Representação da base de dados	4
-----	--	---

Lista de Abreviaturas e Siglas

PDS	Processamento Digital de Sinais
MIDI	Interface Digital de Instrumentos Musicais – do inglês <i>Musical Instrument Digital Interface</i>
DTFT	Transformada de Fourier em Tempo Discreto – do inglês <i>Discret-Time Fourier Transform</i>
DFT	Transformada Discreta de Fourier – do inglês <i>Discret Fourier Transform</i>
FFT	Transformada Rápida de Fourier – do inglês <i>Fast Fourier Transform</i>
STFT	Transformada de Fourier de Tempo Curto – do inglês <i>Short-Time Fourier Transform</i>

Lista de Símbolos

Símbolos Matemáticos

f_s	Frequência de amostragem
f_0	Frequência fundamental
Hz	Hertz - Unidade do Sistema Internacional (SI) para frequência
bpm	Batidas por minutos

Sumário

1	Introdução	1
1.1	Organização do Trabalho	2
1.2	Objetivo	2
2	Fundamentação Teórica	3
2.1	Aprendizado de Máquina	3
2.1.1	Aprendizado Supervisionado	4
2.1.2	Normalização e One-Hot Encoding	5
2.1.3	K-Nearest Neighbors (KNN)	6
2.2	Redes Neurais Artificiais	7
2.2.1	Inspiração Biológica e Perceptron	7
2.2.2	Perceptron multicamadas	9
2.2.3	Camada Softmax	10
2.2.4	Função de Perda(Loss Function)	10
2.2.5	Otimizadores	11
2.2.6	Hiperparâmetros de uma Rede Neural	11
2.2.6.1	Número de camadas e neurônios em cada camada	12
2.2.6.2	Inicialização dos pesos	12
2.2.6.3	Função de Ativação	12
2.2.6.4	Mini-batches	13
2.2.6.5	Taxa de aprendizado	13
2.2.6.6	Número de épocas	13
2.2.7	Regularização	13
2.2.8	Redes Convolutivas	14
2.2.8.1	Camada de Convolução	16

2.2.8.2	Camada Pooling	17
2.2.8.3	Normalização em lote	18
2.2.8.4	Arquitetura de uma Rede Convolutiva	18
2.2.9	Redes Residuais	19
2.3	Séries Temporais	20
2.3.1	Repositório UCR	20
2.3.2	Aplicações de Séries Temporais	21
2.3.3	Medidas de similaridade	21
2.4	Classificação de Séries Temporais	24
2.4.1	Classificação de series temporais com aprendizado profundo	26
2.4.1.1	Modelos Generativos	26
2.4.1.2	Modelos Discriminativos	27
3	Sistema Proposto	28
3.0.1	Conjunto de dados	28
3.0.2	Métricas	28
3.0.3	Metodologia	28
3.1	Visão geral	28
4	Experimentos	29
4.1	Métricas	29
4.2	Objetivos	29
4.3	Resultados Obtidos	29
5	Conclusão	30
	Referências Bibliográficas	31

Capítulo 1

Introdução

Séries temporais estão presentes em diversos fenômenos ao nosso redor. Tanto em atividades humanas, como em processos naturais, elas têm sido alvo de pesquisas em áreas como clima, economia, medicina, transporte. Além disso, em diversos domínios de pesquisa, dados tipicamente não temporais, como figuras geométricas, a face, mão ou postura corporal de um indivíduo, são representados e analisados como séries temporais. Com o aumento do volume de dados gerados e o avanço nas tecnologias de sensores, houve crescimento no tamanho e na complexidade dos fluxos de dados de séries temporais. Assim, a importância e o impacto das técnicas de análise e modelagem de séries temporais continuam a crescer.

No contexto da mineração de dados, uma das tarefas mais importantes é a classificação, cujo objetivo é associar cada instância de um conjunto de dados a uma pré-definida no problema. Os problemas de Classificação de Séries Temporais (*TSC - Time Series Classification*) são diferentes pois a ordem das amostras são relevantes e tem correlação entre si. Diversos modelos clássicos de classificação tratam cada atributo de forma independente.

Para *TSC*, a abordagem mais popular e tradicional é o k-vizinhos mais próximos(KNN) combinado com a distância *Dynamic Time Warping* (DTW). A DTW faz um mapeamento entre pontos semelhantes de duas séries temporais e calcula a distância entre as séries conforme esse mapeamento. Na prática, isto torna a DTW invariante no tempo. Uma extensa avaliação experimental [12] mostrou que o 1-NN DTW é uma referência forte que não pode ser superada por muitos algoritmos propostos. Se vencida, a vantagem muitas vezes não é tão significativa, comparada com a dificuldade da implementação ou do cálculo.

Em 2015 centenas de métodos para classificação de séries temporais foram propostos [12]. Um fator que contribuiu para esse aumento, foi a expansão do repositório UCR *Time Series Classification* para 85 conjuntos de dados de diversos domínios diferentes. O repositório contribuiu para aumentar a qualidade da avaliação de novos algoritmos TSC.

Redes neurais profundas estão sendo progressivamente introduzidas no problema da classificação de séries temporais, com resultados promissores [52] [49] [53]. "PESQUISAR SOBRE APRENDIZADO PROFUNDO COM TSC NOS 3 ARTIGOS"!!!

O objetivo deste trabalho é avaliar diferentes arquiteturas de aprendizagem profunda para classificação de séries temporais com base no repositório da UCR.

Dentro dessa realidade, o presente projeto aborda o desenvolvimento de um sistema de detecção de f_0 para áudios musicais monofônicos, através de processamento digital de sinais, com o intuito de mapear os áudios, fornecendo as frequências fundamentais soadas em cada instante de tempo e identificando os fenômenos aos quais um sistema nesse contexto é submetido. A detecção de f_0 será realizada por meio da Transformada de Fourier de Curto Tempo (STFT), considerando f_0 como a frequência de maior amplitude dentro de cada janela. Foi construída uma base de dados com áudios de 4 tipos de instrumentos musicais para realização de experimentos, visando validar as detecções realizadas pelo sistema. Este trabalho propõe-se ainda a avaliar a metodologia adotada por meio desses experimentos realizados.

1.1 Organização do Trabalho

1.2 Objetivo

Capítulo 2

Fundamentação Teórica

Este capítulo aborda, de forma sucinta, os fundamentos teóricos que foram aplicados no planejamento e implementação deste projeto.

2.1 Aprendizado de Máquina

Aprendizado de máquina(AM) é uma área da inteligência artificial cujo objetivo é o desenvolvimento de técnicas computacionais sobre o aprendizado, bem como a construção de sistemas capazes de adquirir conhecimento de forma automática. Um sistema baseado em aprendizado trabalha por meio de experiências acumuladas e de soluções bem-sucedidas de problemas anteriores [13]. Normalmente, algoritmos de aprendizado utilizam experiências anteriores, denominadas conjunto de treino, para auxiliar no processo de tomada de decisão.

Existem três diferentes tipos de aprendizado: supervisionado, não-supervisionado e semi-supervisionado. A diferença entre esses tipos de aprendizado é se o método utiliza ou não utiliza o rótulo de treino. No aprendizado supervisionado, esse rótulo é conhecido, enquanto que no aprendizado não-supervisionado os exemplos não vistos. Já no aprendizado semi-supervisionado, o conjunto de treinamento consiste de uns poucos exemplos rotulados e muitos não rotulados[14]. O escopo deste trabalho se encontra no aprendizado supervisionado.

	A_1	A_2	\dots	A_M	Classe(Y)
E_1	x_{11}	x_{12}	\vdots	x_{1M}	y_1
E_2	x_{21}	x_{22}	\vdots	x_{2M}	y_2
\vdots	\vdots	\vdots	\ddots	\vdots	\vdots
E_N	x_{N1}	x_{N2}	\vdots	x_{NM}	y_N

Tabela 2.1: Representação da base de dados

2.1.1 Aprendizado Supervisionado

O objetivo do aprendizado supervisionado é construir um modelo que consegue fazer predição através de instâncias de uma base de dados rotuladas. Cada instância é representada por um conjunto de características. Na Tabela 2.1 é mostrada a estrutura de uma base. Neste exemplo cada vetor $E_i = [x_{i1}, \dots, x_{iM}]$ se refere a classe y_i .

A idéia da aprendizagem supervisionada é o conseguir encontrar uma função desconhecida f (função conceito) tal que $y = f(\mathbf{x})$, onde o vetor \mathbf{x} são os atributos de uma instância específica. Na prática, a função f deve conseguir prever o valor correto y_i de uma instância E_i não vista. Normalmente, o número de exemplos da base de dados não é suficiente para descrever a função conceito. Nesse caso, o classificador é visto como uma hipótese h que aproxima f , ou seja, $h(x) \approx f(x)$. Caso os valores dos rótulos y_1, y_2, \dots, y_N sejam numéricos o problema é denominado de *regressão*, caso sejam valores categóricos o problema é denominado de *classificação*.

De maneira geral, a base de dados é dividida em dois conjuntos: conjunto de *treino* e conjunto de *teste*. O conjunto de treinamento é utilizado para ajustar o classificador. Como dito anteriormente, o classificador é uma hipótese da função conceito f , logo, é fundamental que o conjunto de treinamento tenha uma distribuição o mais semelhante possível do conjunto original. O conjunto de teste é utilizado para avaliar o modelo construído. Idealmente, esse conjunto não deve ter exemplos em comum com o conjunto de treinamento.

Em alguns casos, pode ser necessário utilizar um conjunto de *validação*, extraído do conjunto de treinamento, para realizar ajustes no modelo construído pelo algoritmo de aprendizado. Logo tem-se três conjuntos: *treino*, *validação* e *teste*. O treino é utilizado para aprendizagem do algoritmo. O modelo é avaliado através do conjunto de validação. É feita uma alteração dos parâmetros do classificador e outro treinamento é realizado.

O intuito é melhorar o desempenho do modelo através desses "ajustes". Dessa maneira os exemplos de validação são indiretamente "vistos" durante o processo de aprendizado, o que obriga que esses exemplos sejam diferentes dos exemplos de testes.

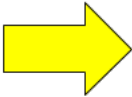
2.1.2 Normalização e One-Hot Encoding

Os algoritmos de aprendizagem de máquina aprendem através dos dados. Dados do mundo real apresentam valores que estão em distintas faixas. A fim de evitar que algum atributo predomine sobre outro ou que inclua alguma ponderação indesejada ao induzir um modelo de AM, é comum fazer uma normalização dos valores de cada atributo. Um forma de normalizar os dados é mostrada na Equação 2.1:

$$x_{ij} = \frac{x_{ij} - \bar{x}}{\sigma_j} \quad (2.1)$$

onde \bar{x} representa a média do atributo e σ_j representa o desvio padrão.

Os algoritmos de AM geralmente possuem como entrada e saída valores numéricos, portanto é necessário converter os valores categóricos da base de dados para valores numéricos. A codificação one-hot é um processo que converte rótulos em vetores binários como mostrado na Figura 2.1. Uma vantagem dessa codificação é que não cria uma "ordem" numérica nos dados. Essa ordem poderia interferir na indução do classificador, podendo dar maior importância para valores maiores, o que não faz sentido para variáveis categóricas.



Color	Red	Yellow	Green
Red	1	0	0
Red	1	0	0
Yellow	0	1	0
Green	0	0	1
Yellow	0	0	1

Figura 2.1: Codificação One-hot

2.1.3 K-Nearest Neighbors (KNN)

Um classificador bastante popular é o K-Nearest Neighbors (K-Vizinhos mais próximos). O KNN utiliza os próprios dados de treinamento como modelo de classificação. Para classificar uma instância de teste, procura-se entre os dados de treinamento os K mais próximos da instância de teste. Por fim, verifica-se qual a classe predominante desses K dados de treinamento, e instância de teste é classificada com essa mesma classe. A cada nova exemplo a ser classificado faz-se uma varredura nos dados de treinamento, o que provoca um grande esforço computacional

O princípio do classificador k-NN é a “regra dos vizinhos mais próximos”. A hipótese é que, dado um conjunto de exemplos distribuídos sobre o espaço de dados X , a “vizinhança” de um exemplo $x \in X$ estabelecida por uma função de distância apropriada tende a ser ocupada por exemplos que pertencem à mesma classes que x [15], como ilustrado na Figura 2.2. Desse modo a informação fornecida pelos exemplos conhecidos que são mais similares a x .

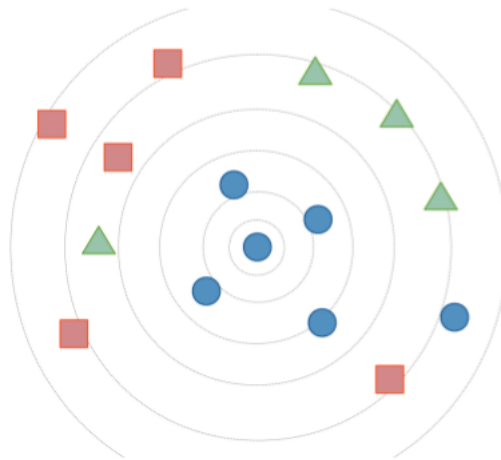


Figura 2.2: Princípio dos k-vizinhos mais próximos

Dois parâmetros importantes devem ser determinados para execução do k -NN:

- **Medida de similaridade:** É a função que compara as instâncias de treino com a instância que se quer classificar. Na seção 2.3.3 será mostrada duas funções de similaridade
- **Número de vizinhos:** Quantidade de vizinhos que será considerada para classificar. Por exemplo, se $k = 1$, então o exemplo é classificado com a mesma classe do exemplo mais próximo segundo a medida de similaridade utilizada. Se $k > 1$,

então são consideradas as classes dos k exemplos mais próximos para realizar a classificação.

2.2 Redes Neurais Artificiais

Redes Neurais Artificiais são modelos computacionais que buscam simular o processamento de informação pelo cérebro humano. Elas são compostas por unidades simples de processamento, os neurônios, que se unem por meio de conexões sinápticas [16]. Cada conexão, além de ser altamente especializada, é responsável pelo envio de sinais de um neurônio para outro. Segundo (Haykin 2009[17]), os neurônios e suas conexões podem ser implementados utilizando-se de componentes eletrônicos ou via simulação programada em computador.

2.2.1 Inspiração Biológica e Perceptron

Um bloco básico de uma rede neural tem algumas semelhanças com um neurônio biológico. O neurônio biológico é uma célula formada por três seções com funções específicas e complementares: *corpo*, *dendritos* e *axônio*. Os dendritos captam os estímulos recebidos em um determinado período de tempo e os transmitem ao corpo do neurônio, onde são processados. Quando tais estímulos atingirem determinado limite, o corpo da célula envia um novo impulso que se propaga pelo axônio e é transmitido às células vizinhas por meio de sinapses. Este processo pode se repetir em várias camadas de neurônios. Como resultado, a informação de entrada é processada, podendo levar o cérebro a comandar reações físicas. [18]. A figura 2.3 ilustra de forma simplificada as partes de um neurônio.

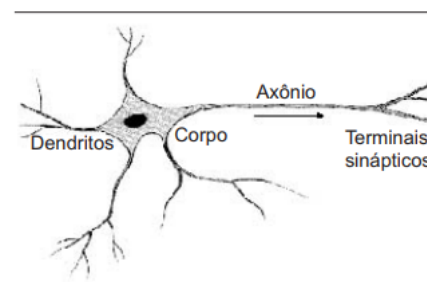


Figura 2.3: Representação simplificada de um neurônio

O modelo de um neurônio artificial é apresentado na Figura ???. Este modelo é com-

posto por três elementos:

- Um conjunto de entradas (x_1, x_2, \dots, x_n) que são multiplicadas por um conjunto de pesos (p_1, p_2, \dots, p_n) ;
- Um somador (Σ) para acumular os sinais de entrada;
- Uma função de ativação que (φ) limita o intervalo permissível de amplitude do sinal de saída (y) a um valor fixo.

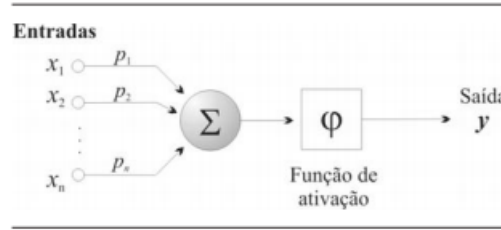


Figura 2.4: Modelo matemático de um neurônio

Esse modelo foi proposto por *McCulloch and Pitts* em 1943 [19] e é conhecido como *perceptron*. A função de ativação (φ) tem a seguinte definição:

$$\varphi = \begin{cases} 1 & \text{if } \sum_{i=1}^n p_i x_i^k \geq T, \\ 0 & \text{if } \sum_{i=1}^n p_i x_i^k < T \end{cases} \quad (2.2)$$

Os valores dos pesos podem ser positivos ou negativos e eles refletem se aquela conexão é inibitória ou excitatória. Um valor positivo ou negativo reflete a importância da respectiva entrada para o processamento. Frequentemente é adicionado um *viés* b na entrada da função de ativação. A forma geral do modelo é descrito como:

$$y(k) = \varphi\left(\sum_{i=1}^n p_i(k)x_i(k) + b(k)\right) \quad (2.3)$$

Os pesos sinápticos do Perceptron podem ser adaptados empregando um processo de aprendizado com um número finito de iterações. A aprendizagem é conduzida pela regra de correção de erro conhecida como algoritmo de convergência do Perceptron. Esse algoritmo visa encontrar um vetor de pesos w tal que as duas igualdades da função degrau sejam satisfeitas (*Lippmann*, 1987 [20])

2.2.2 Perceptron multicamadas

O Perceptron multicamadas (*Multi-Layer Perceptron - MLP*) é uma generalização da rede perceptron. Novas camadas são adicionadas o que possibilita a solução de problemas que não sejam linearmente separáveis. O vetor de entradas \mathbf{x} passa pela camada inicial, cujos valores de saída são ligados a camada seguinte. Esse processo se repete até chegar na última camada. (Figura 2.5) Pode-se arranjar a rede em várias camadas, tornando-a profunda e capaz de aprender relações cada vez mais complexas.

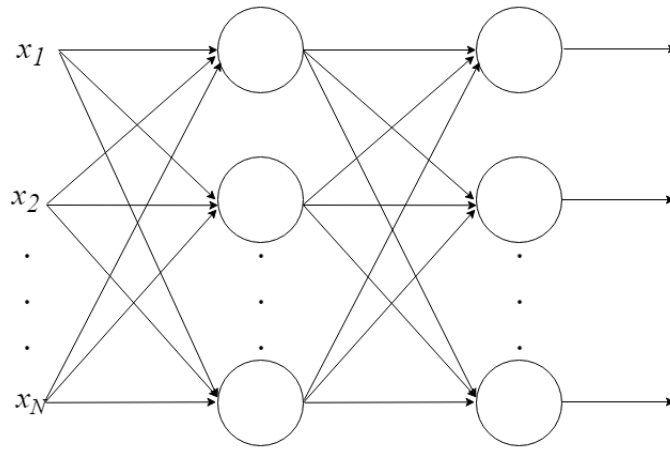


Figura 2.5: Multilayer Perceptron. Cada círculo representa um neurônio mostrado anteriormente

Em 1986, *Rumelhart, Hinton e Williams* [21] desenvolveram o algoritmo *backpropagation*, que utiliza o gradiente descendente para treinar uma MLP. Este método é composto pelas fases *forward* e *backward*. O objetivo do *backpropagation* é otimizar os pesos para que a rede neural possa aprender a mapear corretamente as entradas para as saídas. A primeira fase é a “propagação adiante” (*forward*), onde as entradas inseridas na rede se propagam entre as camadas, uma a uma, até a produção das respectivas saídas, portanto a função dessa fase é gerar uma resposta considerando as entradas e os respectivos pesos sinápticos, os quais permanecem inalterados.

Na fase *backward* é onde a aprendizagem dos pesos é realizada. Esse aprendizado se dá através da otimização (minimização) da função *loss*, a qual determina a qualidade da classificação do dado de entrada. Essa otimização é realizada através de um método chamado *Gradiente Descendente* que busca a minimização da função *loss* ao alterar os pesos na direção de maior declive do gradiente. O gradiente é calculado na última camada e então é retro-propagado para as camadas intermediárias anteriores que contribuem

diretamente para a formação da saída. Cada elemento da camada intermediária recebe é responsável apenas por uma porção do gradiente total. Este processo se repete, camada por camada, até que cada elemento tenha a sua parcela de gradiente para o gradiente total. Baseado no gradiente, é feita uma alteração dos valores dos pesos e bias de modo que a rede aprenda os padrões do conjunto de treinamento.

2.2.3 Camada Softmax

Como explicado em *GoodFellow, Bengio, e Courville*(2016 [22]), a camada softmax é utilizada como um classificador na camada de saída e tem como objetivo representar a probabilidade de cada classe para cada valor de entrada.

Pela Equação 3.5 nota-se que os valores de saída da camada softmax estão entre 0 e 1, e que a soma de de todas as saídas é igual a 1. Desta forma, cada neurônio de saída representa representa a probabilidade da entrada pertencer a uma determinada classe.

$$\text{softmax}(z_i) = \frac{\exp(z_i)}{\sum_j \exp(z_j)} \quad (2.4)$$

2.2.4 Função de Perda(Loss Function)

A função de perda compara a saída da rede para um exemplo de treinamento com o rótulo verdadeiro. Uma função de perda comum é o erro médio quadrático dado pela Equação 2.5

$$MSE = \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{N} \quad (2.5)$$

Quando a saída da rede neural está sendo tratada como uma distribuição de probabilidade é comum utilizar entropia cruzada *cross-entropy*. Ela é usada para quantificar a distâncias entre duas distribuições de probabilidade. Em redes neurais a entropia cruzada comparada a distribuição do que o modelo prediz com a previsão. Ela é definida como:

$$L = - \sum_i y_i \log(z_i) \quad (2.6)$$

Para problemas de classificação a entropia cruzada é mais adequada.

2.2.5 Otimizadores

A finalidade dos otimizadores é minimizar o valor da função de perda alterando nos pesos da redes. Existem atualmente diversos métodos para alcançar esse fim, esses métodos utilizam de alguns hiperparâmetros para calibrar a forma com que os pesos devem ser atualizados. Dentre os mais utilizados se destacam:

- **Stochastic Gradient Descent (SGD) com Nesterov Momentum** é um método onde, em uma alusão à mecânica clássica, a velocidade de descida do resultado da função de perda na superfície multidimensional determinada pelo gradiente é calculada e usada para escalar o tamanho do passo que será dado em direção ao ponto mínimo. Utilizando essa velocidade e o valor do gradiente na posição atual é possível prever a próxima posição dessa partícula virtual onde então é calculado o gradiente da nova posição. Os parâmetros são então atualizados baseados nesse segundo gradiente e em um hiperparâmetro denominado taxa de aprendizado (learning rate). [23]
- **Adagrad** é um método adaptativo de taxa de aprendizado proposto originalmente por [24] onde a taxa de aprendizado é normalizada pela raiz da soma dos gradientes de cada parâmetro ao quadrado. Isso tem o efeito de reduzir a taxa de aprendizado dos pesos que recebem gradientes altos e ao mesmo tempo aumentar a taxa de aprendizado de pesos que recebem atualizações infrequentes. Um hiperparâmetro de suavização também é usado para inibir a divisão por zero.
- **Adam** é um método derivado de um outro método chamado RMSprop que basicamente ajusta o método Adagrad para que a taxa de aprendizado não diminua agressivamente. A contribuição do método Adam é adicionar um momento na atualização (semelhante ao Nesterov Momentum) e suavizar os ruídos do gradiente antes de fazer essa operação. Ele herda do RMSprop adição de uma taxa de decaimento na soma dos gradientes de cada parâmetro que reduz a agressividade de quanto a taxa de aprendizado é reduzida a cada passo. [25]

2.2.6 Hiperparâmetros de uma Rede Neural

A maioria dos algoritmos de AM envolvem "hiperparâmetros", que são variáveis definidas para o algoritmo antes do treinamento com o objetivo de otimizá-lo. Definir os

valores de hiperparâmetros pode ser visto como uma seleção de modelos, ou seja, escolher qual modelo usar do conjunto hipotético de modelos possíveis. Em redes neurais os hiperparâmetros determinam a estrutura da rede (Ex: Número de neurônios em uma camada) ou como a rede será treinada (Ex: Taxa de aprendizado)

2.2.6.1 Número de camadas e neurônios em cada camada

Em problemas de classificação geralmente as redes neurais tem: camada de entrada tem o mesmo número de neurônios do tamanho do vetor de características, camadas escondidas e a camada de saída que tem o mesmo número de classes do problema. Conforme o teorema da aproximação universal [26], uma rede neural com uma única camada oculta com um número finito de neurônios é capaz de aproximar qualquer função contínua de \mathbb{R}^n . Isto quer dizer que teoricamente é possível aproximar qualquer

2.2.6.2 Inicialização dos pesos

Idealmente, pode ser melhor usar diferentes estratégias de inicialização dos pesos conforme a função de ativação aplicada em cada camada. A distribuição uniforme é comumente utilizada. (*)

2.2.6.3 Função de Ativação

As funções de ativações são usadas para introduzir não-lineariedade nos modelos, que permite que as redes aprendam superfícies de decisões mais complexas. Dentre as funções de ativações mais utilizadas estão a Sigmoid e ReLU (*Rectified Linear Units*). A vantagem da ReLU é que ela evita o problema de dissipação do gradiente.

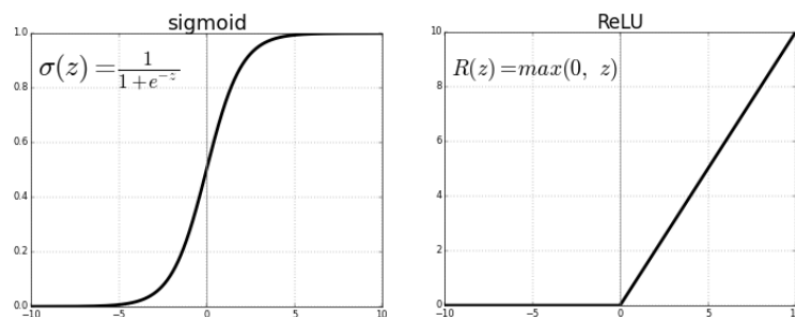


Figura 2.6: Softmax vs ReLU. Imagem tirada do site *Towards Data Science* [1]

2.2.6.4 Mini-batches

O treinamento geralmente é realizado em *batches* também conhecidos como *mini-batches* que são subconjuntos do treino. As razões de utilizar batches e não o treino inteiro são: diminuição do tempo de treinamento; caso o conjunto de treino seja suficientemente grande, uma amostra desse conjunto pode representar de forma *boa* o conjunto original. A tamanho dessa amostra é um hiperparâmetro da rede, sendo geralmente aceito que o treinamento com *batches* maiores resulta numa melhor performance.

2.2.6.5 Taxa de aprendizado

A taxa de aprendizado é um parâmetro constante no intervalo de $[0,1]$ que interfere na convergência do aprendizado. Ela determina o quão “rápido” as atualizações dos pesos irão em direção do gradiente. Se a taxa de aprendizado for muito pequena, o modelo convergirá muito lentamente; Se a taxa de aprendizado for muito grande, o modelo irá divergir.

2.2.6.6 Número de épocas

Número de épocas é a quantidade de vezes que o conjunto de treino é passado pela rede. O recomendável é aumentar o número de épocas até que a acurácia do conjunto de validação diminuir, mesmo quando a acurácia do treino aumentar (*overfitting*).

2.2.7 Regularização

Além do *tuning* dos hiperparâmetros existe outras maneiras para aumentar a performance da rede e torná-la mais robusta em sua capacidade de generalização. Um exemplo é a adição de um termo de regularização na função de perda. Esse termo é introduzido para regularizar os valores dos pesos buscando distribuir melhor o aprendizado por todos os neurônios. A função provavelmente mais usada para este fim é a da regularização L2, onde para cada parâmetro da rede o termo $\frac{1}{2}\lambda\omega^2$ é adicionado ao resultado da função de perda, ω representando o peso em questão, λ sendo a força da regularização e o fator $\frac{1}{2}$ é multiplicado para simplificar o gradiente da função.

Essa regularização penaliza neurônios com pesos muito altos levando a uma distribuição das atualizações para os neurônios com pesos menores. Ao introduzir essa função

é necessário lembrar que os pesos também devem ser atualizados pelo gradiente da função de regularização, o que é um passo geralmente realizado após a atualização quanto ao gradiente da função loss.

A introdução da função de regularização ajuda a combater o overfitting e juntamente com ela uma outra técnica que é muito utilizada para esse fim é o *Dropout* [2]. O Dropout é uma técnica que condiciona à existência de um neurônio na fase de treinamento em função de uma probabilidade predefinida por um hiperparâmetro. Esse método essencialmente desativa neurônios com uma probabilidade p para que a rede seja capaz de otimizar a função loss mesmo sem a presença de todos os seus neurônios.

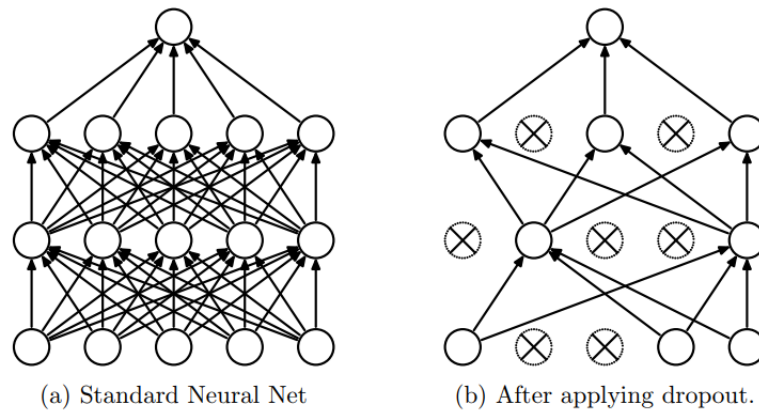


Figura 2.7: Imagem ilustrando o efeito do Dropout em uma rede neural onde os neurônios marcados foram desativados, forçando a rede não depender deles, evitando o *overfitting* [2].

2.2.8 Redes Convolutivas

As redes convolutivas foram desenvolvidas tomando como inspiração o córtex visual de animais, que contém agrupamentos de células que detectam luz em pequena escala que são chamadas de campo receptivos. O campo receptivo de um único neurônio sensorial é a região específica da retina em que a presença de um estímulo causa a ativação dele. Os neurônios sensoriais que tem campos receptivos semelhantes se sobrepõem. Na figura 2.8 é mostrada uma representação de um neurônio e o seu campo receptivo.

Tais células atuam como filtros locais sobre o espaço de entrada e quanto mais complexa a célula, maior o seu campo receptivo. As redes convolutivas seguem esse mesmo princípio de campos receptivos, o que ajuda a diminuir o custo computacional de treinamento. Esse tipo de rede recebe esse nome devido a sua camada mais predominante, a

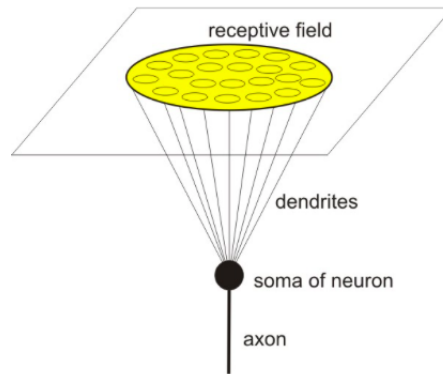


Figura 2.8: Campo receptivo de um neurônio. Fonte: [3]

camada de convolução.

Haykin [27] divide a estrutura das redes convolutivas em três objetivos principais:

- Extração de características: Cada neurônio recebe sinais de entrada de uma determinada janela da camada anterior, possibilitando a extração de características locais. Isto faz com que a posição exata de cada característica (ou pixel, no caso de uma imagem), seja irrelevante, desde que sua posição em relação às características vizinhas seja mantida.
- Mapa de características: As camadas são compostas por diversos mapas de características (*feature maps*) que são regiões onde os neurônios compartilham os mesmos pesos. Esses pesos são chamados de filtros ou kernels, e dão robustez ao modelo, fazendo com que ele seja capaz de lidar com variações de distorção, rotação e translação na imagem. O compartilhamento dos pesos também possibilita uma redução drástica no número de parâmetros a serem otimizados.
- Subamostragem: Após cada camada de convolução é aplicada uma camada de subamostragem (*subsampling*), que é uma coleta de amostras de cada mapa de característica. Essas amostragens podem ser realizadas através da média, maior ou menor valor das amostras. O que produz uma sumarização do mapa de características

A principal aplicação das CNNs é para processamento de informações visuais em particular imagens, pois a convolução permite filtrar as imagens considerando a sua informação visual. Diferente da MLP que vetoriza as imagens a CNN leva em consideração a estrutura espacial entre os pixels vizinhos em uma determinada imagem através da operação de convolução.

A seguir é apresentado os tipos de camadas que geralmente são usadas para redes convolutivas

2.2.8.1 Camada de Convolução

É na camada de convolução que cada característica específica em alguma posição espacial da entrada é detectada. Os parâmetros dessa camada são distribuídos entre os filtros que tem a mesma ideia dos campos receptivos comentados anteriormente. Os filtros são parâmetros ajustáveis do modelo que devem aprender uma característica específica. A convolução ocorre durante a fase *forward*, onde cada um desses filtros são convoluídos com a entrada. Em aprendizagem de máquina os dados são discretos o que acaba simplificando a operação de convolução. O cálculo se torna em um produto entre os elementos de entrada sobrepostos pelos filtros deslocados em toda entrada.

Alguns hiperparâmetros são importantes na configuração de tal camada[28]:

- Tamanho do campo receptivo(*kernel size*): Para o caso de imagens é o tamanho bidimensional do campo receptivo que será convoluído com a entrada, e pode ter valores como 3x3, 5x5, 16x16. É o mesmo para todos os neurônios da camada. Para o caso em 1-D esse hiperparâmetro é apenas um inteiro.
- Profundidade: É o número de kernels que serão estimulados pela mesma parte do campo receptivo. Cada um desses kernels é responsável por detectar uma característica por toda entrada, constituindo então um conjunto de mapas de características. A Figura 2.9 ilustra uma camada de convolução com profundidade cinco.

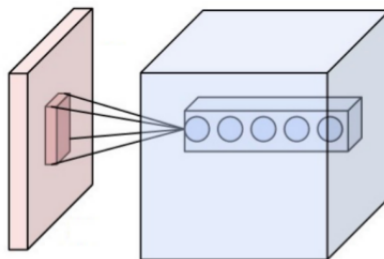


Figura 2.9: Disposição dos neurônios na camada de convolução e seu campo receptivo. Fonte: [4]

- Stride: É o parâmetro que determina o espaçamento da sobreposição dos campos receptivos. A Figura 2.10 mostra que valores diferentes de stride podem alterar o

campo receptivo de tais neurônios e o número total dos neurônios de cada fatia da profundidade. Esse parâmetro é decisivo na capacidade de identificar os padrões visuais da entrada, sendo responsável por definir quão refinada será a identificação do padrão mesmo que trasladado, onde o aumento desse refinamento é inversamente proporcional ao desempenho da camada, por conta do maior número de neurônios da mesma.

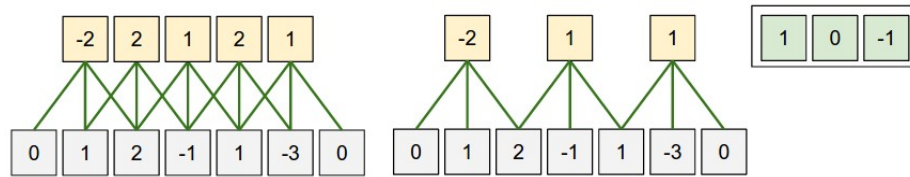


Figura 2.10: Neurônios espaçados por stride 1 e 2 das esquerda para a direita respectivamente. Separados à direita estão os pesos compartilhados entre todos os neurônios. Fonte: [5]

- Zero-padding: Determina qual tamanho do preenchimento com zeros que será adicionado às bordas da entrada. Isso muitas vezes é utilizado para adequar o tamanho da entrada ao stride escolhido, como pode ser visto na Figura 2.10 onde o tamanho da entrada (representada unidimensionalmente) é cinco e o zero-padding é um.

2.2.8.2 Camada Pooling

A camada de pooling faz uma subamostragem da entrada, onde cada valor de saída é encontrado pela a média ou valor máximo de uma janela definida. Na figura 2.11 é mostrado o funcionamento de cada pooling. A entrada é uma imagem 4x4 e a saída é obtida através de uma janela 2x2. O valor de saída é calculado conforme o tipo de pooling. Para o pooling médio é calculada médio da janela. Para o caso máximo é encontrado o maior valor da janela.

Estas pequenas amostras de acordo com [22], melhora o poder computacional da rede pois as camadas posteriores vão ficando cada vez menores o que resulta em uma diminuição do uso de memória. Outra vantagem do pooling é que ele ajuda na invariância a translação da entrada.

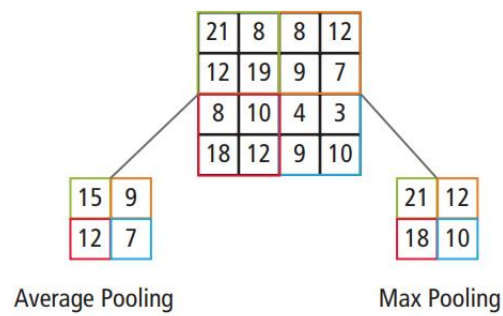


Figura 2.11: Representação de max-pooling e average-pooling. Fonte: [5]

2.2.8.3 Normalização em lote

Conforme explicado na Seção 2.1.2 o conjunto de dados da entrada é normalizado para evitar problemas de comparação devido a diferentes escalas dos dados. Ocorre que durante a passagem dos exemplos normalizados através das camadas da rede, esses valores são novamente transformados (por meio das funções de ativações), podendo fazer com que os dados de entrada de algumas camadas ocultas fiquem desnormalizados novamente. Esse problema é conhecido como *internal covariate shift* e apresenta maior gravidade quanto mais profunda for a rede a ser treinada. Esse problema aumenta o tempo de treinamento porque implica na definição de uma taxa de aprendizagem pequena, além de propiciar a dissipação dos gradientes.

A normalização em lote (do inglês batch normalization) é um mecanismo proposto recentemente para resolver o problema acima, e que consiste em normalizar os dados fornecidos a cada camada oculta. A normalização é aplicada em cada mini-lote, para aumentar a eficiência durante a aplicação da transformação. Esse método tem um efeito regularizador, tendo substituído o uso de regularização L1 e L2 e em alguns casos tornando desnecessário o uso de dropout.

2.2.8.4 Arquitetura de uma Rede Convolutiva

Após todo o processo de convolução realizado nas camadas anteriores, em muitas estruturas de CNNs é usado antes de sua saída, pelo menos uma camada completamente preenchida com as ativações anteriores, como mostrado na Figura 2.12, nomeada de "Fully Connected". Nesta imagem é mostrada uma arquitetura de rede convolutiva para reconhecimento de dígitos de 0 a 9.

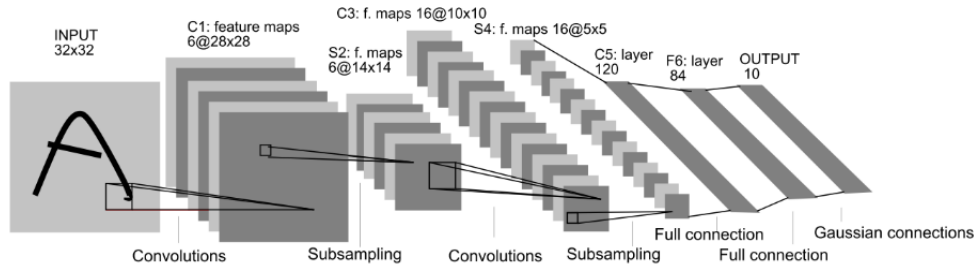


Figura 2.12: Arquitetura LeNet-5. Fonte: [6]

As redes convolutivas são treinadas utilizando *backpropagation* através de um método de otimização conforme explicado na Seção 2.2.5. É necessário definir os hiperparâmetros como número de filtros, tamanho do kernel, função de ativação, número de camadas entre outros. Isso pode ser feito através de métodos de validação cruzada como *10-fold* ou *leave-one-out*.

2.2.9 Redes Residuais

As Redes Residuais, também conhecidas como *ResNets*[7], se diferenciam das CNNs convencionais pois incorporam uma técnica para evitar a degradação de dados nas camadas mais profundas da rede e o desaparecimento do gradiente durante o *backpropagation*. Uma *ResNet* possui camadas residuais que são compostas, cada uma, de duas camadas convolutivas e duas funções de ativação ReLU, conforme demonstra a Figura 2.13. A camada residual difere das demais camadas convolutivas pois sua saída é a soma da saída da segunda camada convolutiva com a entrada da camada residual. Essa estrutura possibilita que camadas mais profundas recebam diretamente dados das camadas mais superficiais, sem a necessidade de passar por camadas convolutivas intermediárias.

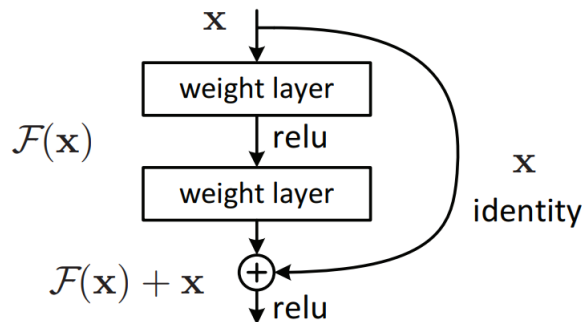


Figura 2.13: Bloco Residual. Fonte: [7]

2.3 Séries Temporais

Série temporal é uma sequência de observações de um fenômeno ao longo do tempo e que possui intervalos de tempos iguais para cada par de observações. A ordem das amostras é crucial, pois há uma dependência entre os dados e uma alteração da ordem pode modificar o significado dos dados. As séries temporais estão em diversas áreas do conhecimento como Economia (preços diários de ações, taxa mensal de desemprego, produção industrial), Medicina (eletrocardiograma, eletroencefalograma), Epidemiologia (número mensal de novos casos de meningite), Meteorologia (precipitação pluviométrica, temperatura diária, velocidade do vento).

Uma série temporal pode ser definida como:

$$X(t) = (x_1, x_2, \dots, x_n) \quad (2.7)$$

onde x_n representa uma observação no instante t , n o número de observações e $X(t)$ a função que descreve a série temporal em termos de t . Caso a série seja constituída de uma observação em cada instante de tempo ela é chamada de *univariada*. Caso a série foi obtida por uma coleta simultânea de dois ou mais fenômenos ela é chamada de *multivariada*.

As séries temporais estão em diversas áreas do conhecimento como Economia (preços diários de ações, taxa mensal de desemprego, produção industrial), Medicina (eletrocardiograma, eletroencefalograma), Epidemiologia (número mensal de novos casos de meningite), Meteorologia (precipitação pluviométrica, temperatura diária, velocidade do vento).

Embora séries temporais tenham sido estudadas por mais de 100 anos na estatística, muitos de seus métodos não são escaláveis para análise de grandes volumes de séries temporais, o que aumenta o interesse no uso de métodos de mineração de dados e aprendizagem de máquina. [29]

2.3.1 Repositório UCR

O conjunto de dados utilizados neste trabalho foi o repositório UCR [30]. Com nome oficial de “*UCR Time Series Classification Archive*”, esse repositório foi criado em 2002 pela Universidade da Califórnia em Riverside (daí vem o acrônimo UCR) inicialmente com dezesseis conjunto de dados de séries temporais. Atualmente possui 128 conjuntos

de dados (com uma atualização recente em 2018) e reúnem vários domínios de aplicação. Esses conjuntos de dados possuem partições padrão de treinamento e teste, facilitando a realização dos experimentos e a comparação de resultados. Tornou-se um importante recurso na comunidade de mineração de dados em séries temporais, com mais de mil artigos publicados que utilizam pelo menos um conjunto de dados do repositório [31].

2.3.2 Aplicações de Séries Temporais

As principais tarefas de interesse para mineração de séries temporais são sintetizadas a seguir [29]:

- *Classificação*: dado um conjunto de séries temporais não rotuladas S , treinar um classificador f , utilizando um conjunto de séries temporais previamente rotuladas, para associar S a uma das classes do problema.
- *Agrupamento*: o objetivo é distribuir as séries temporais em grupos, de modo que séries temporais de um mesmo grupo são mais similares entre si, do que a séries temporais de outros grupos, de acordo com uma função de similaridade;
- *Previsão*: dada uma série temporal, o objetivo é prever observações futuras.
- *Detecção de Motivos*: o objetivo é encontrar uma ou mais subsequências que aparecem frequentemente na séries;
- *Detecção de Anomalias*: dada uma série temporal S considerada normal e uma série temporal T , o objetivo é encontrar qualquer parte de T que seja considerada anômala, inesperada ou de interesse.

2.3.3 Medidas de similaridade

Em mineração de séries temporais, para tarefas como classificação, agrupamento e indexação, é importante definir a função de distância usada para calcular a similaridade entre pares de séries temporais. A seguir, são apresentadas as principais duas funções de distância para séries temporais propostas na literatura.

Uma das principais funções de distância utilizadas para análise de séries temporais, e em diversos domínios de aplicação, é a Distância Euclidiana. Dada duas séries temporais

de valores reais $A = \{a_1, a_2, \dots, a_m\}$ e $B = \{b_1, b_2, \dots, b_m\}$, a distância Euclidiana é definida como:

$$ED(A, B) = \sqrt{\sum_{i=1}^m (a_i - b_i)^2} \quad (2.8)$$

Uma desvantagem da distância Euclidiana é a limitação no cálculo de distância entre séries temporais com mesmo comprimento (número de observações). Na Figura 2.14, são mostradas as associações feitas entre cada par de observações de duas séries temporais para o cálculo da distância Euclidiana.

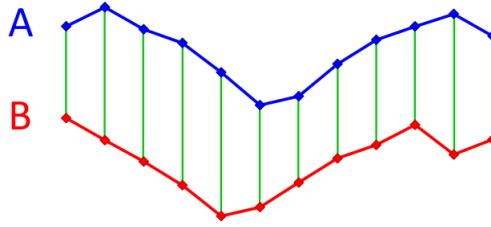


Figura 2.14: Distância Euclidiana. Fonte: [8]

É possível notar que as séries temporais mostradas na Figura 2.14 apresentam comportamento muito similar, no entanto a série temporal B está deslocada uma unidade de tempo em relação à série temporal A. O uso da distância Euclidiana não permite o reconhecimento desse deslocamento, pois são calculadas as diferenças entre pares de observações sempre no mesmo instante de tempo.

Nesse caso, uma alternativa é a Dynamic Time Warping (DTW)[32], que permite realizar o alinhamento entre duas séries temporais e, assim, reconhecer a similaridade entre pares de séries, mesmo se elas apresentarem comprimentos diferentes ou se elas não estiverem alinhadas em relação à escala temporal.

Sejam duas séries temporais $A = \{a_1, a_2, \dots, a_m\}$ e $B = \{b_1, b_2, \dots, b_n\}$ de tamanhos m e n respectivamente, constrói-se uma matriz m -por- n onde o elemento (i, j) da matriz contém a distância $d(a_i, b_j)$ entre dois os pontos a_i e b_j (geralmente é utilizada a Distância Euclidiana). Cada elemento (i, j) corresponde ao alinhamento entre os pontos a_i e b_j . Isso é ilustrado na Figura 2.15.

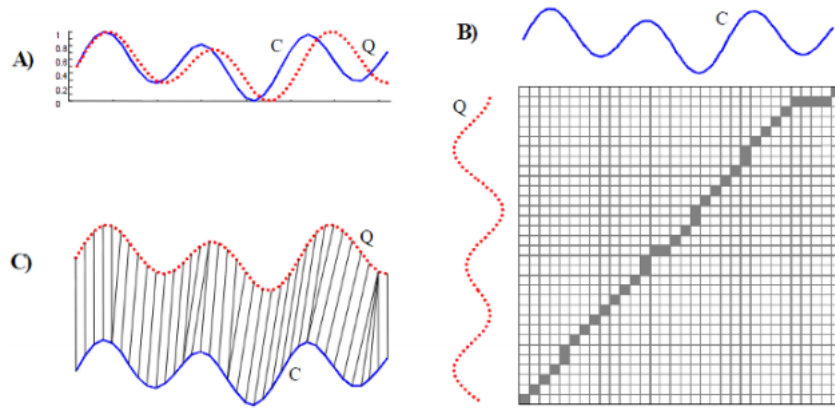


Figura 2.15: A) Duas sequências Q e C são semelhantes, porém fora de fase. B) Para alinhar as sequências, construímos a matriz e procuramos pelo melhor caminho. C) O alinhamento resultado. Fonte: [9]

Um caminho W é um conjunto de elementos da matriz que define um mapeamento entre A e B . Um elemento k de W é definido como $w_k = (i, j)_k$. O caminhamento W é sujeito a algumas restrições:

- **Limites:** $w_1 = (1, 1)$ e $w_k = (m, n)$, isso requer que o caminho comece e termine em cantos diagonalmente opostos a matriz.
- **Continuidade:** Sendo $w_k = (a, b)$ então $w_{k-1} = (a, b)$ onde $a - a \leq 1$ e $b - b \leq 1$, isso restringe os possíveis passos do caminhamento para células que sejam adjacentes (incluindo diagonalmente adjacentes).
- **Monotonicidade:** Sendo $w_k = (a, b)$ então $w_{k-1} = (a, b)$ onde $a - a \geq 0$ e $b - b \geq 0$, isso força os pontos de W a serem monotonicamente espaçados no tempo.

Existem vários caminhos que satisfazem as condições acima, entretanto no cálculo de DTW, o objetivo é resolver o problema:

$$DTW(A, B) = \min \sqrt{\sum_{k=1}^K w_k} \quad (2.9)$$

De forma intuitiva, o DTW está interessado em obter o caminho que minimiza as distâncias entre as séries. Este caminho pode ser encontrado utilizando programação

dinâmica para buscar os próximos elementos do conjunto W . Assim, a distância cumulativa definida pela programação dinâmica $\gamma(i, j)$ e o mínimo das distâncias cumulativas das células adjacentes podem ser expressadas como:

$$\gamma(i, j) = d(a_i, b_j) + \min\{\gamma(i-1, j-1), \gamma(i-1, j), \gamma(i, j-1)\} \quad (2.10)$$

A Distância Euclidiana entre duas sequências pode ser vista como um caso especial de DTW. Nesse caso, os elementos k de W são restringidos à $w_k = (i, j)_k$ onde $i = j = k$. Também esse caso especial só pode ser aplicado se as duas sequências possuem o mesmo tamanho, ou seja $n = m$.

Na Figura 2.16, é exibido o resultado da aplicação da DTW para calcular a distância entre as séries temporais. Pode-se notar que a DTW realiza o alinhamento entre as duas séries temporais, de modo a reconhecer a similaridade entre seus comportamentos, mesmo havendo um deslocamento na escala temporal.

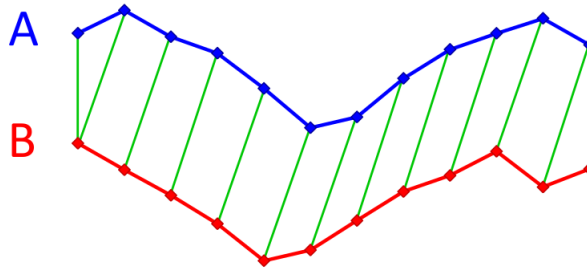


Figura 2.16: A similaridade entre as séries temporais A e B é reconhecida pela DTW. [8]

2.4 Classificação de Séries Temporais

Problemas de diversas áreas podem ser modelados como problemas de classificação de séries temporais, tais como: reconhecimento de fala, detecção de movimentos e detecção de arritmias em eletrocardiogramas. Estes são alguns dos exemplos comumente encontrados.

Alta dimensionalidade, alta correlação entre as *features* e a grande quantidade de ruído são algumas das características que tornam as séries temporais desafiadoras aos algoritmos clássicos de classificação. Essa dificuldade as torna um problema interessante que tem sido explorado pela literatura.

Métodos de classificação de séries temporais podem ser divididos em duas categorias [33]:

- *Baseados em atributos:* extraem atributos das séries temporais para depois usar métodos de aprendizagem de máquina clássica.
- *Baseados em instâncias:* classificam as instâncias de teste com base na similaridade com as instâncias de treino.

Grande parte das contribuições tem sido voltadas para a métodos baseados em atributos com definição de medidas de similaridade entre as séries temporais. [34].

Diversos estudos evidenciam que a distância DTW combinada com um classificador 1-NN é uma das abordagens mais competitivas em tarefas de classificação [35] [36] [37]. Contudo existem algumas desvantagens. O 1-NN é um classificador do tipo “preguiçoso” pois ele precisa verificar toda a base de treino para classificar uma instância. O cálculo da DTW tem complexidade $O(nm)$, onde n e m são os tamanhos das séries temporais. Ou seja, para cada passo do 1-NN é feito um cálculo da DTW, o que torna a classificação de toda a base de teste muito custosa, especialmente se o tamanho das séries forem grandes.

Existem abordagens que tem como objetivo diminuir o tempo computacional da DTW. É esperado que o caminho ótimo não esteja muito distante da diagonal da matriz de distâncias. Tomando isso como pressuposto é possível definir uma banda de *warping* onde a busca pelo caminho ótimo é limitada. Na prática, isso impede o alinhamento entre pares de observações muito distantes no eixo do tempo, o que também evita um problema conhecido como *warping patológico* [38]. Na figura 2.17 é mostrada dois tipos de bandas, a janela de Sakoe-Chiba e o paralelogramo de Itakura.

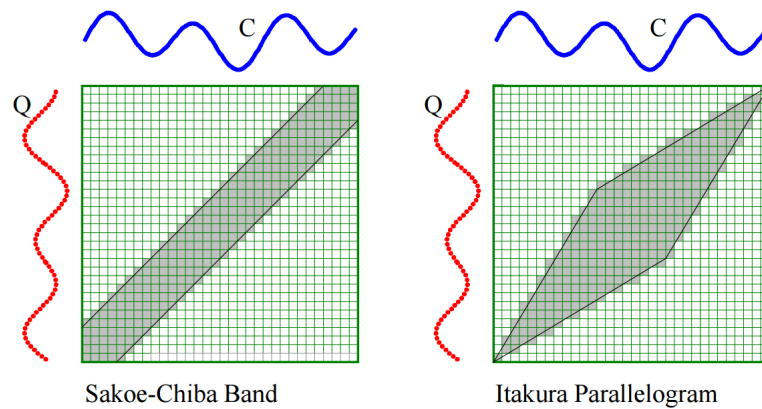


Figura 2.17: Visualização da janela de Sakoe-Chiba e do paralelogramo de Itakura. Fonte: [10]

Dentre as abordagens mais sofisticadas, os melhores resultados são alcançados por *Bag-of-SFA-Symbols* (BOSS)[39] e *Collective of transform-based ensembles*(COTE) [40]. O primeiro método é um procedimento de transformação de séries temporais inspirado em processamento de texto, enquanto o segundo é um conjunto que agrega os resultados de 35 classificadores que processam quatro tipos de representação de séries temporais.

2.4.1 Classificação de series temporais com aprendizado profundo

As abordagens de redes profundas para classificação de séries temporais podem ser divididas em duas categorias[11]: modelos generativos e discriminativos, conforme mostrado na Figura 2.18

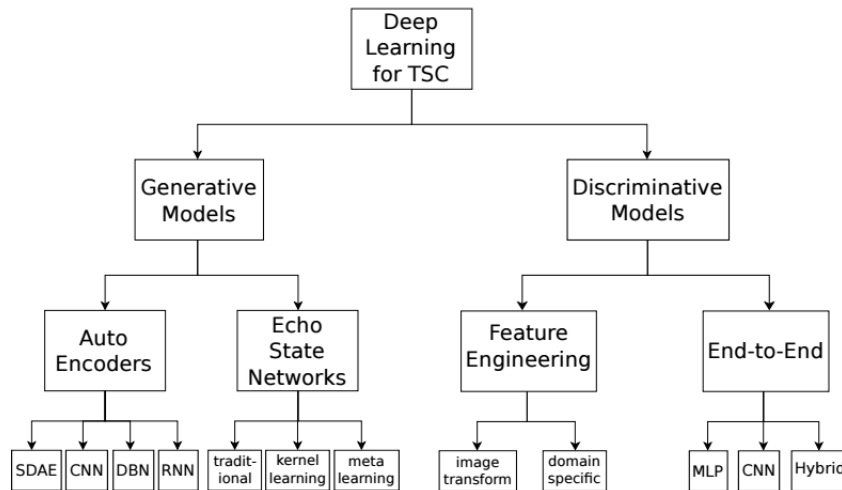


Figura 2.18: Visão geral de diferentes abordagens de aprendizado profundo para classificação de séries temporais. Fonte: [11]

2.4.1.1 Modelos Generativos

Modelos generativos geralmente passam por uma etapa não-supervisionada que precede o de treinamento do classificador. O objetivo dos modelos generativos é encontrar uma boa representação das séries temporais antes de treinar um classificador. Um exemplo de modelo generativo é o autocodificador que procura aprender uma representação comprimida da entrada, chamada espaço latente. Neste caso o autocodificador funciona como um extrator de características automático.

Modelos generativos baseado em redes convolutivas são propostos em [41] [42] onde

os autores usam a operação de deconvolução e *upsampling* que ajudam a reconstruir series temporais multivariadas. Em [43] [44], os autores utilizam autocodificadores com redes recorrentes para gerar séries temporais em um espaço latente e por fim treinar um classificador, como *SVM* ou *Random Forest*, que prevê a classe de uma determinada série temporal de entrada.

Outra abordagem em modelos generativos são as Redes com Estado de Eco (*Echo State Networks - ESN*) que são utilizadas em [45] [46]. As ESN são redes neurais compostas por: uma camada de entrada; uma camada interna, denominada reservatório de dinâmicas (RD); e uma camada de saída com característica linear; [47]. Em classificação de séries temporais, as ESN são utilizadas para transformar a entrada em um espaço conhecido como *reservoir*, que é utilizado para classificação.

2.4.1.2 Modelos Discriminativos

Modelos discriminativos aprendem a mapear a entrada bruta, ou através de Engenharia de Características (*Feature Engineering*), de uma série temporal e gera uma distribuição de probabilidade sobre as classes do problema em questão na camada de saída.

Engenharia de Características é o processo de transformar dados brutos em características que melhor representam o problema, resultando em um aumento na precisão do modelo [48]. As abordagens com engenharia de características e aprendizado profundo, envolvendo séries temporais, mais utilizadas são aquelas que transformam a entrada em imagens tais como Campos Gramianos, Gráficos de Recorrência e Campos de Transição de Markov [49] [50] [51].

Outra forma de modelos discriminativos são aqueles em que são fornecidos os dados temporais brutos para a rede profunda. Em [52], o autor propõe uma CNN com múltiplos canais para séries multivariadas, onde cada canal recebe como entrada uma única dimensão da série multivariada e aprende as características de maneira separada. Em [53], o autor projetou uma MLP, FCN (Fully Convolutional Network) e uma ResNet. Em [54], o autor propõe uma arquitetura combinada de LSTM-FCN (*Long Short-Term Memory*) e conseguiu atingir o estado da arte em termos de classificação de séries temporais com redes profundas.

Capítulo 3

Sistema Proposto

Neste capítulo é descrito as arquiteturas escolhidas

3.0.1 Conjunto de dados

Conforme apresentado anteriormente, a classificação de séries temporais tem aplicações nas mais diferentes áreas. Dessa forma, faz-se necessário avaliar e analisar o comportamento do método proposto em dados de diferentes tipos de aplicações. Para os experimentos conduzidos neste capítulo, foram selecionados 11 conjuntos de dados do UCR Time Series Classification Archive[30].

3.0.2 Métricas

Para comparar os m

3.0.3 Metodologia

3.1 Visão geral

Capítulo 4

Experimentos

4.1 Métricas

4.2 Objetivos

4.3 Resultados Obtidos

Capítulo 5

Conclusão

Referências Bibliográficas

- [1] Towards Data Science. *Activations Functions: Neural Networks*. 2017. [Online; acessado em 18 de Dezembro, 2013]. Disponível em: <<https://towardsdatascience.com/activation-functions-neural-networks-1cbd9f8d91d6>>.
- [2] SRIVASTAVA, N. et al. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, v. 15, p. 1929–1958, 2014. Disponível em: <<http://jmlr.org/papers/v15/srivastava14a.html>>.
- [3] neuroclusterbrain.com. Neuron model rf-psth. 2017. [Online; acessado em 08 de Janeiro, 2019]. Disponível em: <<http://neuroclusterbrain.com>>.
- [4] APHEX34. Input volume connected to a convolutional layer. 2015. [Online; acessado em 08 de Janeiro, 2019]. Disponível em: <https://commons.wikimedia.org/wiki/File:Conv_layer.png>.
- [5] KARPATY, A. *Illustration of spatial arrangement*. 2015. [Online; acessado em 17 de Janeiro, 2019]. Disponível em: <<https://github.com/cs231n/cs231n.github.io/blob/master/assets/cnn/stride.jpeg>>.
- [6] LECUN, Y. et al. Gradient-based learning applied to document recognition. In: *Proceedings of the IEEE*. [S.l.: s.n.], 1998. p. 2278–2324.
- [7] HE, K. et al. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. Disponível em: <<http://arxiv.org/abs/1512.03385>>.
- [8] AMARAL, B. F.
Classificação semissupervisionada de séries temporais extraídas de imagens de satélite
— Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São

- Carlos, 2016. Dissertação de Mestrado em Ciências de Computação e Matemática Computacional. Disponível em: <http://www.teses.usp.br/teses/disponiveis/55/55134/tde-18112016-105621/>.
- [9] KEOGH, E.; RATANAMAHATANA, C. A. Exact indexing of dynamic time warping. *Knowledge and Information Systems*, v. 7, n. 3, p. 358–386, Mar 2005. ISSN 0219-3116. Disponível em: <https://doi.org/10.1007/s10115-004-0154-9>.
- [10] RATANAMAHATANA, C.; KEOGH, E. J. Making time-series classification more accurate using learned constraints. In: *SDM*. [S.l.: s.n.], 2004.
- [11] FAWAZ, H. I. et al. Deep learning for time series classification: a review. *CoRR*, abs/1809.04356, 2018. Disponível em: <http://arxiv.org/abs/1809.04356>.
- [12] BAGNALL, A. et al. The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery*, v. 31, n. 3, p. 606–660, May 2017. ISSN 1573-756X. Disponível em: <https://doi.org/10.1007/s10618-016-0483-9>.
- [13] MONARD, M.; BARANAUSKAS, J. *Conceitos sobre aprendizado de máquinas. Em Sistemas Inteligentes: Fundamentos e Aplicações*. 1º. ed. [S.l.]: Editora Manole, 2003.
- [14] CHAPELLE, O.; SCHÖLKOPF, B.; ZIEN, A. *Semi-Supervised Learning*. 1º. ed. [S.l.]: MIT Press, Cambridge, 2006. 12 – 13 p.
- [15] COVER, T.; HART, P. Nearest neighbor pattern classification. *Information Theory, IEEE Transaction on*, 1967.
- [16] ZHANG, G.; PATUWO, B. E.; HU, M. Y. Forecasting with artificial neural networks: The state of art. *International Journal of Forecasting*, 1998.
- [17] HAYKIN, S. S. *Neural networks and learning machines*. 3º. ed. Upper Saddle River, United States of America: Prentice Hall.
- [18] FERNEDA, E. Redes neurais e sua aplicação em sistemas de recuperação de informação. *Ciência da Informação*, v. 35, n. 1, 2006. ISSN 1518-8353. Disponível em: <http://revista.ibict.br/ciinf/article/view/1149>.

- [19] MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 1943. ISSN 1522-9602. Disponível em: <<https://doi.org/10.1007/BF02478259>>.
- [20] LIPPMANN, R. An introduction to computing with neural nets. *IEEE ASSP Magazine*, v. 4, n. 2, 1987.
- [21] RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning representations by back-propagating errors. *Nature*, Nature Publishing Group, v. 323, p. 533–, out. 1986. Disponível em: <<http://dx.doi.org/10.1038/323533a0>>.
- [22] GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. [S.l.]: MIT Press, 2016. <http://www.deeplearningbook.org>.
- [23] BENGIO, Y.; BOULANGER-LEWANDOWSKI, N.; PASCANU, R. Advances in optimizing recurrent networks. *CoRR*, abs/1212.0901, 2012. Disponível em: <<http://arxiv.org/abs/1212.0901>>.
- [24] DUCHI, J. C.; HAZAN, E.; SINGER, Y. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, v. 12, p. 2121–2159, 07 2011.
- [25] KINGMA, D. P.; BA, J. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. Disponível em: <<http://arxiv.org/abs/1412.6980>>.
- [26] CYBENKO, G. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems. 1989 Springer-Verlag New York Inc*, p. 303–314, 1989.
- [27] HAYKIN, S. S. *Neural networks and learning machines*. Third. Upper Saddle River, NJ: Pearson Education, 2009.
- [28] GUEDES, A. B. *Reconhecimento de Gestos usando Redes Convolucionadas*. 49 p. — UnB - Universidade de Brasília, Brasília, 2017.
- [29] MAIMON, O.; ROKACH, L. *Data Mining and Knowledge Discovery Handbook*. Berlin, Heidelberg: Springer-Verlag, 2005. ISBN 0387244352, 9780387244358.

- [30] CHEN, Y. et al. *The UCR Time Series Classification Archive*. 2015. www.cs.ucr.edu/~eamonn/time_series_data/.
- [31] DAU, H. A. et al. Ucr time series archive. *CoRR*, abs/1810.07758, 2018.
- [32] BERNDT, D. J.; CLIFFORD, J. Using dynamic time warping to find patterns in time series. In: *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*. AAAI Press, 1994. (AAAIWS'94), p. 359–370. Disponível em: <<http://dl.acm.org/citation.cfm?id=3000850.3000887>>.
- [33] DENG, H. et al. A time series forest for classification and feature extraction. *Inf. Sci.*, Elsevier Science Inc., New York, NY, USA, v. 239, p. 142–153, ago. 2013. ISSN 0020-0255. Disponível em: <<https://doi.org/10.1016/j.ins.2013.02.030>>.
- [34] KEOGH, E.; KASETTY, S. On the need for time series data mining benchmarks: A survey and empirical demonstration. *Data Min. Knowl. Discov.*, Kluwer Academic Publishers, Hingham, MA, USA, v. 7, n. 4, p. 349–371, out. 2003. ISSN 1384-5810. Disponível em: <<https://doi.org/10.1023/A:1024988512476>>.
- [35] WANG, X. et al. Experimental comparison of representation methods and distance measures for time series data. *Data Mining and Knowledge Discovery*, v. 26, n. 2, p. 275–309, Mar 2013. Disponível em: <<https://doi.org/10.1007/s10618-012-0250-5>>.
- [36] LINES, J.; BAGNALL, A. Time series classification with ensembles of elastic distance measures. *Data Mining and Knowledge Discovery*, v. 29, n. 3, p. 565–592, May 2015. Disponível em: <<https://doi.org/10.1007/s10618-014-0361-2>>.
- [37] Giusti, R.; Batista, G. E. A. P. A. An empirical comparison of dissimilarity measures for time series classification. In: *2013 Brazilian Conference on Intelligent Systems*. [S.l.: s.n.], 2013. p. 82–88.
- [38] RATANAMAHATANA, C. A.; KEOGH, E. Everything you know about dynamic time warping is wrong. In: CITESEER. *Third Workshop on Mining Temporal and Sequential Data*. [S.l.], 2004.
- [39] SCHÄFER, P. The boss is concerned with time series classification in the presence of noise. *Data Mining and Knowledge Discovery*, v. 29, n. 6, p. 1505–1530, Nov 2015. ISSN 1573-756X. Disponível em: <<https://doi.org/10.1007/s10618-014-0377-7>>.

- [40] BAGNALL, A. et al. Time-series classification with cote: The collective of transformation-based ensembles. *IEEE Transactions on Knowledge and Data Engineering*, v. 27, p. 1–1, 09 2015.
- [41] ZHIGUANG et al. Representation learning with deconvolution for multivariate time series classification and visualization. *CoRR*, abs/1610.07258, 2016. Disponível em: <<http://arxiv.org/abs/1610.07258>>.
- [42] MITTELMAN, R. Time-series modeling with undecimated fully convolutional neural networks. 08 2015. Disponível em: <<https://arxiv.org/abs/1508.00317>>.
- [43] MEHDIYEV, N. et al. Time series classification using deep learning for process planning: A case from the process industry. *Procedia Computer Science*, v. 114, p. 242 – 249, 2017. ISSN 1877-0509. Complex Adaptive Systems Conference with Theme: Engineering Cyber Physical Systems, CAS October 30 – November 1, 2017, Chicago, Illinois, USA. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S1877050917318707>>.
- [44] MALHOTRA, P. et al. Timenet: Pre-trained deep recurrent neural network for time series classification. *CoRR*, abs/1706.08838, 2017. Disponível em: <<http://arxiv.org/abs/1706.08838>>.
- [45] ASWOLINSKIY, W.; REINHART, R. F.; STEIL, J. Time series classification in reservoir- and model-space. *Neural Processing Letters*, v. 48, n. 2, p. 789–809, Oct 2018. ISSN 1573-773X. Disponível em: <<https://doi.org/10.1007/s11063-017-9765-5>>.
- [46] BIANCHI, F. M. et al. Reservoir computing approaches for representation and classification of multivariate time series. 03 2018.
- [47] SANTOS, M. S.
Classificadores neurais de treinamento rápido aplicados na identificação online de evento no detector de atlas — Universidade Federal da Bahia, Salvador, 2018. Dissertação de Mestrado em Engenharia Elétrica. Disponível em: <http://www.ppgee.eng.ufba.br/teses/c2125147ac78f314a43ec5d461e07f73.pdf>.
- [48] J. Brownlee. *Discover Feature Engineering, How to Engineer Features and How to Get Good at It*. 2014. [Online; acessado em 11 de março, 2019]. Dis-

- ponível em: <<https://machinelearningmastery.com/discover-feature-engineering-how-to-engineer-features-and-how-to-get-good-at-it/>>.
- [49] WANG, Z.; OATES, T. Imaging time-series to improve classification and imputation. In: *Proceedings of the 24th International Conference on Artificial Intelligence*. AAAI Press, 2015. (IJCAI'15), p. 3939–3945. ISBN 978-1-57735-738-4. Disponível em: <<http://dl.acm.org/citation.cfm?id=2832747.2832798>>.
- [50] TRIPATHY, R.; ACHARYA, U. R. Use of features from rr-time series and eeg signals for automated classification of sleep stages in deep neural network framework. *Biocybernetics and Biomedical Engineering*, v. 38, n. 4, p. 890 – 902, 2018. ISSN 0208-5216. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S0208521618300949>>.
- [51] WANG, Z.; OATES, T. Spatially encoding temporal correlations to classify temporal data using convolutional neural networks. *CoRR*, abs/1509.07481, 2015. Disponível em: <<http://arxiv.org/abs/1509.07481>>.
- [52] ZHENG, Y. et al. Time series classification using multi-channels deep convolutional neural networks. In: *WAIM*. [S.l.: s.n.], 2014.
- [53] WANG, Z.; YAN, W.; OATES, T. Time series classification from scratch with deep neural networks: A strong baseline. *CoRR*, abs/1611.06455, 2016. Disponível em: <<http://arxiv.org/abs/1611.06455>>.
- [54] Karim, F. et al. Lstm fully convolutional networks for time series classification. *IEEE Access*, v. 6, p. 1662–1669, 2018. ISSN 2169-3536.