UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Câmpus de Ilha Solteira - SP

GLORIA PATRICIA LÓPEZ SEPÚLVEDA

APLICAÇÃO DE INTELIGÊNCIA COMPUTACIONAL NA RESOLUÇÃO DE PROBLEMAS DE SISTEMAS ELÉTRICOS DE POTÊNCIA

Ilha Solteira
2017

GLORIA PATRICIA LÓPEZ SEPÚLVEDA

APLICAÇÃO DE INTELIGÊNCIA COMPUTACIONAL NA RESOLUÇÃO DE PROBLEMAS DE SISTEMAS ELÉTRICOS DE POTÊNCIA

Tese apresentada à Faculdade de Engenharia - UNESP - Câmpus de Ilha Solteira, para obtenção do título de Doutora em Engenharia Elétrica.
Área de conhecimento: Automação.

Prof. Dr. Marcos Julio Rider Flores
Orientador

Ilha Solteira
2017

FICHA CATALOGRÁFICA
Desenvolvido pelo Serviço Técnico de Biblioteca e Documentação

López Sepúlveda, Gloria Patricia.
L925a Aplicação de inteligência computacional na resolução de problemas de sistemas elétricos de potência / Gloria Patricia López Sepúlveda. -- Ilha Solteira: [s.n.], 2017
181 f. : il.

Tese (doutorado) - Universidade Estadual Paulista. Faculdade de Engenharia de Ilha Solteira. Área de conhecimento: Automação, 2017

Orientador: Marcos Julio Rider Flores
Inclui bibliografia

1. Aprendizado de máquina. 2. Árvores de decisão. 3. Carregamento de veículos elétricos. 4. Controle centralizado Volt-VAr. 5. Detecção de fraudes. 6. Inteligência computacional.

Raiane da Silva Santos
Serviço Técnico de Biblioteca e Documentação
Seção Técnica de Referência, Atendimento ao Usuário e Documentação
Supervisora Técnica de Seção
CRB CRB-8/00000

unesp

UNIVERSIDADE ESTADUAL PAULISTA

Câmpus de Ilha Solteira

CERTIFICADO DE APROVAÇÃO

TÍTULO DA TESE: Aplicação de Inteligência Computacional na Resolução de Problemas de Sistemas Elétricos de Potência

AUTORA: GLORIA PATRICIA LOPEZ SEPULVEDA
ORIENTADOR: MARCOS JULIO RIDER FLORES

Aprovada como parte das exigências para obtenção do Título de Doutora em ENGENHARIA ELÉTRICA, área: AUTOMAÇÃO pela Comissão Examinadora:

Prof. Dr. MARCOS JULIO RIDER FLORES
Departamento de Engenharia Elétrica / Universidade Estadual de Campinas - UNICAMP

Prof. Dr. RUBEN AUGUSTO ROMERO LAZARO
Departamento de Engenharia Elétrica / Faculdade de Engenharia de Ilha Solteira

Prof. Dr. JOHN FREDY FRANCO BAQUERO
Departamento de Engenharia Elétrica / Universidade Estadual Paulista Júlio de Mesquita Filho, Câmpus Experimental Rosana

Prof. Dr. ADRIANO BATISTA DE ALMEIDA
Centro de Engenharia e Ciências Exatas / Universidade Estadual do Oeste do Paraná

Prof. Dr. JUAN CARLOS GALVIS MANSO
Departamento de Engenharia Elétrica / Universidade Federal de Ouro Preto

Ilha Solteira, 13 de novembro de 2017

À minha família, em especial ao meu esposo Hugo, à minha filha Valentina, à minha mãe Carmen, à minha vovó Dioselina e à meu vovô Jorge, por todo o amor, apoio, confiança e incentivo em todas as batalhas enfrentadas ao longo da minha vida.

AGRADECIMENTOS

Meus agradecimentos a todos os familiares, amigos, professores e funcionários da FEIS-UNESP, que direta ou indiretamente contribuíram para a realização deste trabalho. Em especial, dedico meus agradecimentos:

- A Deus, por ter me dado força e saúde para chegar até aqui;
- A minha família pelo carinho, apoio e incentivo;
- Ao meu esposo Hugo Andres pelo amor, apoio, confiança e incentivo em todos os momentos;
- Ao Prof. Dr. Marcos Julio Rider Flores, por todo ensinamento, incentivo, confiança, paciência e orientação;
- Ao Prof. Dr. Rubén Romero Lázaro, pelo acompanhamento durante este tempo, sugestões e incentivo;
- Ao Prof. Dr. Fredy Franco pela co-orientação, paciência e todos os ensinamentos;
- Ao Prof. Dr. Carlos Julio Zapata pela ajuda na obtenção de dados reais;
- Ao Engenheiro Julio Gomez por permitir executar os testes desta tese em sistemas reais;
- Aos meus amigos e colegas do laboratório que de forma direta ou indireta me ajudaram;
- A Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes), pela oportunidade e apoio financeiro.

“Quando dois homens imaginam a mesma coisa,

ainda assim cada um tem sua própria ideia.”

Frege

RESUMO

Nesta tese são utilizados algoritmos de Inteligência Computacional para resolver quatro problemas da área de sistemas elétricos de potência, com o intuito de automatizar a tomada de decisões em processos que normalmente são realizados por especialistas humanos ajudados de métodos computacionais clássicos. Nesta tese são utilizados os algoritmos de aprendizado de máquina: árvores de decisão, redes neurais artificiais e máquinas de vetor de suporte, para realizar o processo de aprendizado dos sistemas inteligentes e para realizar a mineração de dados. Estes algoritmos podem ser treinados a partir das medições disponíveis e ações registradas nos centros de controle dos sistemas de potência. Sistemas Inteligentes foram utilizados para realizar: a) o controle centralizado Volt-VAr em modernos sistemas de distribuição de energia elétrica em tempo real usando medições elétricas; b) a detecção de fraudes nas redes de distribuição de energia elétrica realizando um processo de mineração de dados para estabelecer padrões de consumo que levem a possíveis clientes fraudadores; c) a localização de faltas nos sistemas de transmissão de energia elétrica automatizando o processo de localização e ajudando para que uma ação de controle da falta seja realizada de forma rápida e eficiente; e d) a coordenação de carga inteligente de veículos elétricos e dispositivos de armazenamento em tempo real utilizando a tecnologia V2G, nos sistemas de distribuição de energia elétrica a partir de medições elétricas. Para o problema de controle centralizado Volt-VAr os testes foram realizados em um sistema de 42 barras, para o problema de carregamento de veículos elétricos e de dispositivos de armazenamento os testes foram realizados usando um sistema de 34 barras, e os outros dois problemas foram testados com dados reais fornecidos por empresas do setor elétrico colombiano. O software WEKA, versão 3.8.0, foi utilizado para gerenciar os três algoritmos de aprendizado de máquina através do treinamento e validação dos algoritmos Multilayer Perceptron/Backpropagation para as redes neurais artificiais, o J48/C4.5 para as árvores de decisão e o SMO/PoliyKernel para as máquinas de vetor de suporte. Através dos resultados obtidos é possível comprovar o potencial dos Sistemas Inteligentes e da Mineração de Dados no desenvolvimento de algoritmos de automação para os quatro problemas da área de sistemas elétricos de potência.

Palavras-chave: Aprendizado de máquina. Árvores de decisão. Carregamento de veículos elétricos. Controle centralizado Volt-VAr. Detecção de fraudes. Inteligência computacional. Localização de faltas. Máquinas de vetor de suporte. Mineração de dados. Redes neurais artificiais. Sistemas inteligentes.

ABSTRACT

In this thesis Computational Intelligence algorithms are used to solve four problems of the area of power electrical systems, in order to automate decision making in processes that are usually performed by human experts aided by classical computational methods. In this thesis the machine learning algorithms are used: decision trees, artificial neural networks and support vector machines to carry out the learning process of Intelligent Systems and to perform Data Mining. These algorithms are trained from the available measurements and actions recorded in the control centers of the systems. Intelligent Systems were used to perform: a) the centralized control Volt-VAr in modern systems of distribution of electrical energy in real time using electrical measurements; b) detection of fraud in electricity distribution networks by performing a data mining process to establish patterns of consumption that lead to possible fraudulent customers; c) fault location in electric power transmission systems by automating the localization process and helping to ensure that a fault control action is performed quickly and efficiently; and d) coordination of intelligent charging of electric vehicles and storage devices using V2G technology in real-time, in electric power distribution systems using electrical measurements. For the centralized control problem Volt-VAr was tested in 42-node distribution system, for the problem of loading electric vehicles and storage devices the tests were performed using a system of 34-node, and the other two problems were tested with actual data provided by companies in the Colombian electric sector. The WEKA software, version 3.8.0, was used to manage the three machine learning algorithms through the training and validation of the Multilayer Perceptron / Backpropagation algorithms for the artificial neural networks, the J48 / C4.5 for the decision trees and the SMO/PoliyKernel for vector support machines. Through the results obtained it is possible to prove the potential of Intelligent Systems and Data Mining in the development of automation algorithms for the four problems of the area of electrical power systems.

Keywords: Artificial neural networks. Intelligent systems. Centralized control Volt-VAr. Computational intelligence. Data mining. Decision trees. Fault location. Fraud detection. Loading of electric vehicles. Machine learning. Support vector machines.

LISTA DE FIGURAS

Figura 1 Taxonomia das metodologias utilizadas dentro da área de IC. 31

Figura 2 AD e as regiões de decisão no espaço de objetos. 37

Figura 3 Curva de aprendizagem do algoritmo de AD em 100 exemplos gerados aleatoriamente. 37

Figura 4 Modelo matemático simples de um neurônio. 40

Figura 5 (a) Rede neural *feedforward*. (b) Rede neural recorrente. 42

Figura 6 (a) Rede *perceptron* com dois nós de entrada e dois nós de saída. (b) Rede neural com dois nós entradas, uma camada oculta de dois nós e dois nós de saída. 43

Figura 7 Comparação de desempenho dos perceptrons e das ADs. 44

Figura 8 Estrutura de uma RNA. 45

Figura 9 MVS (a) Duas classes de pontos. (b) Separador de margem máxima 50

Figura 10 Hierarquia entre Dado, Informação e Conhecimento. 53

Figura 11 Etapas do Processo de DCBD. 56

Figura 12 Seleção de dados. 57

Figura 13 Redução de dados. 57

Figura 14 Etapas do processo de DCBD. 59

Figura 15 Descrição do processo de MD. 60

Figura 16 Diagrama das variáveis de entradas e saídas do SI. 69

Figura 17 Descrição das variáveis de controle nos diferentes instantes de tempo. . . . 70

Figura 18 Sistema de distribuição de 42 nós. 71

Figura 19 Banco de dados de treinamento. 72

Figura 20 Processo de treinamento realizado entre o MPLIM e o SI. 75

Figura 21 Perdas de energia calculadas e ação de controle definida pelos SIs durante 24 horas. 77
Figura 22 Tensão mínima calculada e ação de controle definida pelos SIs durante 24 horas. 78
Figura 23 Processo de treinamento realizado entre o MPLIM e o SI. 79
Figura 24 Perdas de energia calculadas e ação de controle definida pelo SI durante 24 horas. 81
Figura 25 Tensão mínima calculadas e ação de controle definida pelo SI durante 24 horas. 82
Figura 26 Características das perdas. 86
Figura 27 Percentual de perdas de cada empresa distribuidora em 2014. 88
Figura 28 Percentual de Perdas em Relação à Energia Injetada no Sistema Global das 64 Distribuidoras. 88
Figura 29 Percentual de perdas de cada empresa distribuidora em 2016. 89
Figura 30 Percentual de Perdas em Relação à Energia Injetada no Sistema Global das 63 Distribuidoras. 89
Figura 31 Tipos de fraudes nas redes de DEE. 97
Figura 32 Alteração do medidor. 98
Figura 33 Pré-processamento dos dados. 100
Figura 34 Árvore de Decisão (Atributo CLA_CON). 101
Figura 35 Árvore de Decisão (Atriibuto ANOM). 103
Figura 36 Tipos de faltas nas linhas de transmissão. 113
Figura 37 Causas das faltas nos sistemas de transmissão de energia elétrica. 114
Figura 38 Esquema geral de linha curta. 117
Figura 39 Sinal de corrente. Registro dos relés em uma perturbação real da falta. . . . 119
Figura 40 Sinal de tensão. Registro dos relés em uma perturbação real da falta. 119
Figura 41 Esquema de medição utilizado para obtenção de dados. 120
Figura 42 Diagrama unifilar da rede de transmissão do sistema Colombiano. 121
Figura 43 Modelo dos SIs com suas variáveis de entrada e saída. 122

Figura 44 Processo descritivo do funcionamento dos SIs. 123
Figura 45 Dados filtrados. 125
Figura 46 Comparação do erro médio absoluto. 126
Figura 47 Comparação da raiz do erro médio quadrado. 126
Figura 48 Comparação do erro absoluto relativo. 127
Figura 49 Comparação da raiz do erro relativo ao quadrado. 127
Figura 50 Comparação do tempo de treinamento dado em segundos. 128
Figura 51 Comparação do tempo de validação dado em segundos. 128
Figura 52 Localização de faltas ADs Vs falta simulada. 129
Figura 53 Localização de faltas RNA Vs falta simulada. 129
Figura 54 Localização de faltas RNA Vrs falta simulada. 129
Figura 55 Localização de faltas ADs Vrs falta simulada. 130
Figura 56 Diagrama de abstração de dados do MPLIM para criação da base de dados. . 138
Figura 57 Diagrama das variáveis de entradas e saídas do SI apartir da base de dados. . 139
Figura 58 Diagrama de avaliação de cada SI. 140
Figura 59 Dados que compõem a base de dados inicial. 142
Figura 60 Dados da base de dados apos à aplicação do filtro. 142
Figura 61 RNA MLP com 14 camadas ocultas. 143
Figura 62 RNA MLP com 3 camadas ocultas. 143
Figura 63 Percentagem de classificação para os diferentes algoritmos que compõem os SIs 144
Figura 64 Comparação de tempo de treinamento entre os algoritmos que compõem os SIs 145
Figura 65 Comparação de tempo de validação entre os algoritmos que compõem os SIs 145
Figura 66 Janela inicial do WEKA (GUI Chooser). 163
Figura 67 Interface de linha de comando do WEKA. 164
Figura 68 Ambiente Explorer. 165
Figura 69 Menu do Open URL. 166

Figura 70 Menu do Open DB. 167
Figura 71 Menu do Generate DB. 168
Figura 72 Aplicação de filtros da interface **Explorer**. 169
Figura 73 Interface de classificação. 170
Figura 74 Menu de opções adicionais na interface **Classify**. 171
Figura 75 Ambiente Knowledge Flow. 173
Figura 76 Ambiente Experimenter. 174
Figura 77 Parâmetros a serem calibrados de uma RNA MLP. 176
Figura 78 Parâmetros a serem calibrados de um algoritmo de MVS SMO. 178
Figura 79 Parâmetros a serem calibrados de um algoritmo de MVS SMO. 178
Figura 80 Parâmetros a serem calibrados de um algoritmo de AD J48. 179

LISTA DE TABELAS

Tabela 1 Perdas de energia do sistema obtidas durante uma semana usando as diferentes metodologias 75

Tabela 2 Valores calculados pelo MPLIM e definidos pelo SI das perdas de energia do sistema hora a hora em kWh 76

Tabela 3 Parâmetros utilizados para o treinamento da RNA 79

Tabela 4 Perdas de energia em kWh obtidas durante uma semana usando as diferentes metodologias 80

Tabela 5 Valores calculados pelo MPLIM e definidos pelo SI das perdas do sistema hora a hora em kWh 80

Tabela 6 Percentagem de acertos e DMA por variável para 20% de variação na demanda 83

Tabela 7 DMA dos valores obtidos (5% de variação da demanda) 83

Tabela 8 DMA dos valores obtidos (10% de variação da demanda) 83

Tabela 9 DMA dos valores obtidos (15% de variação da demanda) 83

Tabela 10 DMA dos valores obtidos (20% de variação da demanda) 84

Tabela 11 Percentagem de resultados previstos com valores de DMA entre 0% e 20% 84

Tabela 12 Porcentagem de classificação usando Cross-Validation 84

Tabela 13 Comparação de perdas globais entre os anos 2014 e 2016. 92

Tabela 14 Matriz de Confusão (Atributo OSC). 101

Tabela 15 Matriz de Confusão (Atributo ANOM). 103

Tabela 16 Comparação dos resultados obtidos do algoritmo LibSVM. 105

Tabela 17 Comparação dos resultados obtidos do algoritmo SMO. 106

Tabela 18 Testes com o algoritmo de MVS 107

Tabela 19 Parâmetros do algoritmo de RNA 108

Tabela 20 Testes com o modelo RNA 108
Tabela 21 Dados gerados pelo Digsilent. 120
Tabela 22 Parâmetros da linha de transmissão. 121
Tabela 23 Medidas estatisticas do erro absoluto. 128
Tabela 24 Percentagem de acerto na classificação durante a construção do modelo para diferentes arquiteturas de RNAs MLP 144
Tabela 25 Ferramentas que gerenciam SIs. 162
Tabela 26 Descrição dos parâmetros associados a técnica de RNA. 177
Tabela 27 Descrição dos parâmetros associados a técnica de MVS. 179
Tabela 28 Descrição dos parâmetros associados ao algoritmo J48. 180

LISTA DE ABREVIAÇÕES E SIGLAS

AD	Arvore de Decisão
ABRADEE	Associação Brasileira de Distribuidores de Energia Elétrica
AM	Aprendizado de Máquina
AMPL	*A Modeling Language for Mathematical Programming*
ANOM	Anomalias
ANEEL	Agência Nacional de Energia Elétrica
AR	Alto Risco
BC	Bancos de Capacitores
BR	Baixo Risco
CA	Consumo Alto
CB	Consumo Baixo
CCIVE	Coordenação de Carga Inteligente de Veículos Elétricos
CM	Consumo Médio
CLA_CON	Classificação do Consumo
CSV	*Comma-Separated Values*
DA	Dispositivos de Armazenamento
DCBD	Descoberta do Conhecimento em Bases de Dados
DEE	Distribuição de Energia Elétrica
ECEE	Empresas Concessionárias de Energia Elétrica
ETEE	Empresas de Transmissão de Energia Elétrica
GD	Gerador Distribuido
IA	*Inteligência Artificial*
IDC	Internet Das Coisas
IC	Inteligência Computacional
KDD	*Knowledge Discovery in Databases*
MD	Mineração de Dados
MLP	*MultiLayer Perceptron*
MPLIM	Modelo Matematico de Programação Linear Inteiro Misto
MR	Médio Risco
MVS	Máquinas de Vetor de Suporte
OLTC	*On-Load Tap Changers*
OSC	Oscilação
PCSOBM	Programação Cônica de Segunda Ordem Binário Misto
PNLIM	Programação Não Linear Inteiro Misto

POSD	Planejamento da Operação dos Sistemas de Distribuição
PSO	*Particle Swarm Optimization*
RNA	Redes Neurais Artificiais
RIS	Risco
RT	Reguladores de Tensão
SBC	Sistema Baseado em Conhecimento
SCADA	Supervisão, Controle e Aquisição de Dados
SDEE	Sistemas de Distribuição de Energia Elétrica
SETA	Sistema Especialista para Tratamento de Alarmes
SEP	Sistemas Elétricos de Potência
SI	Sistemas Inteligentes
SMO	*Support Vector Machine*
SR	Sem Risco
STEE	Sistemas de Transmissão de Energia Elétrica
VEs	Veículos Elétricos
VEHs	Veículos Elétricos Híbridos
WEKA	*Waikato Environment for Knoowledge Analysis*

LISTA DE SÍMBOLOS

θ_i	Ângulo de fase na barra i
g_{ij}	Condutância da linha no ramo ij
Y	Conjunto das linhas que podem ou não serem adicionadas no ramo ij
Ω_b	Conjunto de barras
Ω_l^1	Conjunto de caminhos nos quais existem Linhas na configuração base
Ω_l^2	Conjunto de caminhos novos (onde serão adicionadas novos Linhas)
Ω_l^0	Conjunto de linhas existentes na configuração base
Ω_l	Conjunto de ramos
c_{ij}^n	Custo de construção das linhas no ramo ij
d_i	Demanda na barra i
ε_f	Error da condição de factibilidade
ε_o	Error da condição de otimalidade
ε_μ	Error do parâmetro de barreira
γ	Fator de segurança
$\overline{f}_{ij}^0$	Fluxo de potência ativa máximo nos ramos para o conjunto de linhas já existentes
$\overline{f}_{ij}^1$	Fluxo de potência ativa máximo nos ramos para o conjunto de linhas já existentes ou linhas adicionadas em paralelo
$\overline{f}_{ij}^2$	Fluxo de potência ativa máximo nos ramos para o conjunto de linhas correspondentes aos novos caminhos
$\overline{f}_{ij}$	Fluxo de potência ativa máximo permitida no ramo ij para linhas novas
f_{ij}^0	Fluxo de potência ativa nos ramos para o conjunto de linhas já existentes
f_{ij}^1	Fluxo de potência ativa nos ramos para o conjunto de linhas já existentes ou linhas adicionadas em paralelo
f_{ij}^2	Fluxo de potência ativa nos ramos do conjunto de linhas correspondentes aos novos caminhos
f_{ij}	Fluxo de potência ativa no ramo ij para linhas novas
$f_{ij,y}$	Fluxo na linha y do ramo ij
p_i	Geração na barra i
$\overline{p}_i$	Geração máxima na barra i
v	Investimento devido às adições de Linhas no sistema - Função Objetivo
ij	Linha entre as barras i e j
n_{ij}	Número de linhas adicionadas no ramo ij

$\overline{n}_{ij}^{2}$	Número máximo de linhas em caminhos novos
$\overline{n}_{ij}^{1}$	Número máximo de linhas que podem ser adicionadas em paralelo às linhas dos caminhos já existentes
$\overline{n}_{ij}$	Número máximo de Linhas que podem ser adicionados no ramo ij
n_{ij}^{1}	Número de linhas adicionadas em paralelo às linhas já existentes
n_{ij}^{0}	Número de linhas existentes na configuração base no ramo ij
n_{ij}^{2}	Número de linhas novas adicionadas no ramo ij
γ_{ij}	Susceptância nas linhas do ramo ij
γ_{ij}^{0}	Susceptância nas linhas existente do ramo ij
$w_{ij,y}$	Variável binária correspondente à linha y candidata a ser adicionada ou não no ramo ij
x_{ij}	reatância do circuíto ij
q_i	vetor de geração de potência reativa na barra i
$\overline{q_i}$	limite máximo de geração de potência reativa na barra i
$\underline{q_i}$	limite mínimo de geração de potência reativa na barra i
e_i	vetor de demanda de potência reativa na barra i
V_i	magnitude de tensão na barra i
$\overline{V_i}$	limite máximo da magnitude de tensão na barra i
$\underline{V_i}$	limite mínimo da magnitude de tensão na barra i
e_i	vetor de demanda de potência reativa na barra i
s_{ij}^{de}	fluxo de potência aparente (MVA) no ramo ij saindo do terminal
s_{ij}^{para}	fluxo de potência aparente (MVA) no ramo ij chegando no terminal
$\overline{s_{ij}}$	limite de fluxo de potência aparente (MVA) no ramo ij
θij	diferença angular entre as barra i e j
Ω_{bi}	conjunto das barras vizinhas da barra I
g_{ij}	condutância da linha no ramo ij
g_{ij}^{0}	condutância existente da linha no ramo ij
b_{ij}	susceptância da linha no ramo ij
b_{ij}^{sh}	susceptância shunt da linha no ramo ij
b_{i}^{sh}	susceptância shunt na barra i
G_{ij}	matriz de condutância
B_{ij}	matriz de susceptância

SUMÁRIO

1 INTRODUÇÃO 22

1.1 CONTEXTO 24

1.2 OBJETIVOS 27

1.2.1 Objetivo geral 27

1.2.2 Objetivos específicos 27

1.3 CONTRIBUIÇÕES DA TESE 27

1.4 ESTRUTURA DA TESE 28

2 INTELIGÊNCIA COMPUTACIONAL 29

2.1 INTRODUÇÃO 29

2.2 SISTEMAS INTELIGENTES 32

2.3 PARADIGMAS DE APRENDIZAGEM 33

2.3.1 Paradigma simbólico 34

2.3.1.1 Aprendizagem em AD 35

2.3.1.2 Escolha de atributos 38

2.3.1.3 Generalização e super-adaptação 38

2.3.1.4 Aplicabilidade de AD 39

2.3.2 Paradigma conexionista 40

2.3.2.1 Redes Neurais Artificiais 40

2.3.2.2 Estrutura das RNAs 41

2.3.2.3 Rede Perceptron 43

2.3.2.4 Processo de Treinamento 47

2.3.3 Paradigma estatístico 49

2.3.3.1 Máquinas de Vetor de Suporte *49*

2.4 MINERAÇÃO DE DADOS 52

2.4.1 Sistema Baseado em Conhecimento **53**

2.4.2 Sistemas de Descoberta do Conhecimento em Bases de Dados **54**

2.4.3 Etapas do processo de DCBD **55**

2.4.3.1 Etapa de pré-processamento *56*

2.4.3.2 Etapa de Mineração de Dados *59*

2.4.3.3 Etapa de pós-processamento *61*

2.5 CONSIDERAÇÕES FINAIS 62

3 APLICAÇÃO DE SIs NO CONTROLE CENTRALIZADO Volt-VAr EM MODERNOS SDEE **64**

3.1 INTRODUÇÃO 64

3.1.1 Formulação Cônica de Segunda Ordem Binário Misto para o Problema de POSD Radiais **66**

3.2 METODOLOGIA PROPOSTA 68

3.3 CASO DE ESTUDO E RESULTADOS 71

3.3.1 Treinamento dos SIs **74**

3.3.2 Validação dos SIs **79**

3.4 CONCLUSÕES DO CAPÍTULO 84

4 DETECÇÃO DE FRAUDES NAS REDES DE DISTRIBUIÇÃO DE ENERGIA ELÉTRICA UTILIZANDO TÉCNICAS DE MINERAÇÃO DE DADOS **86**

4.1 INTRODUÇÃO 86

4.2 METODOLOGIA PROPOSTA 93

4.2.1 Etapa de Pré-processamento **94**

4.2.2 Etapa de Mineração de Dados **96**

4.2.3 Etapa de Pós-processamento **98**

4.3 CASO DE ESTUDO E RESULTADOS 99

4.3.1 Pré-Processamento dos Dados 99

4.3.2 Mineração de Dados 100

4.3.2.1 Aplicação do Algoritmo de AD J48 101

4.3.2.2 Aplicação do algoritmo de MVS 104

4.3.2.3 Aplicação do algoritmo de RNA 107

4.3.3 Pós-processamento 109

4.4 CONCLUSÕES DO CAPÍTULO 109

5 LOCALIZAÇÃO DE FALTAS EM LINHAS DE TRANSMISSÃO DE ENERGIA ELÉTRICA APLICANDO ALGORITMOS DE INTELIGÊNCIA COMPUTACIONAL 111

5.1 INTRODUÇÃO 111

5.2 MÉTODOS UTILIZADOS PARA LOCALIZAÇÃO DE FALTAS 115

5.3 METODOLOGIA PROPOSTA 118

5.4 CASO DE ESTUDO E RESULTADOS 123

5.5 CONCLUSÕES DO CAPÍTULO 130

6 SISTEMAS INTELIGENTES PARA COORDENAÇÃO DE CARGA ÓTIMA DE VEÍCULOS ELÉTRICOS E DISPOSITIVOS DE ARMAZENAMENTO CONSIDERANDO A TECNOLOGIA V2G 132

6.1 INTRODUÇÃO 132

6.1.1 Métodos utilizados para coordenação de carga de Veículos Elétricos 134

6.2 METODOLOGIA PROPOSTA 137

6.3 CASO DE ESTUDO E RESULTADOS 140

6.3.1 Treinamento e validação dos SIs 142

6.4 CONCLUSÕES DO CAPÍTULO 146

7 CONCLUSÕES E TRABALHOS FUTUROS 147

REFERÊNCIAS 151

APÊNDICE A - WEKA **161**

A.1 INTRODUÇÃO 161

A.1.1 Escolha de plataforma de aprendizagem de máquina **162**

A.2 DESCRIÇÃO DA FERRAMENTA WEKA 163

A.2.1 Simple CLI (Command Line Interface) **164**

A.2.2 Explorer **164**

A.2.2.1 Preprocess ***166***

A.2.2.2 Aplicação de Filtros ***168***

A.2.2.3 Classify ***169***

A.2.3 Knowledge Flow **172**

A.2.4 Experimenter **173**

A.2.5 Classificação com RNAs e MVSs no WEKA **175**

1 INTRODUÇÃO

O crescimento mundial da demanda de energia elétrica, tem provocado preocupações a nível mundial. Tal crescimento, basicamente é devido ao desenvolvimiento económico. No Brasil, esta preocupação cresce, na medida em que o sistema elétrico brasileiro vem operando com níveis de geração próximos às demandas dos usuários. Considerando os investimentos estatais e privados, a preocupação permanece, pois o nível de crescimento atual supera os investimentos sendo realizados e já os compromete quando finalizados. As entidades reguladoras do mercado de energia elétrica preocupadas com cenários como este, exigem das empresas geradoras de energia a garantia do fornecimento para manter o consumo crescente. Neste contexto é de fundamental importância a otimização dos processos relacionados com a transmissão e a distribuição de energia elétrica, e para tal é necessário o uso de metodologias interdisciplinares que auxiliem na otimização e na automação destes processos (BUSH, 2013).

Considerando a preocupação das empresas concessionárias dos sistemas de distribuição de energia elétrica (SDEE) e dos sistemas de transmissão de energia elétrica (STEE) em serem mais flexíveis aos avanços tecnológicos, ao aumento nos preços dos combustíveis e ao interesse da conservação do meio ambiente, incentiva-se o desenvolvimento da geração distribuída, sistemas de armazenamento de energia, programas de resposta à demanda e tecnologias de medição sincronizada, como componentes essenciais para atingir as redes inteligentes ou também conhecidas como *smart grids* (WERBOS, 2011).

A denominação de rede inteligente gera a noção de "inteligência" embutida dentro da rede de energia elétrica. Uma alternativa esta relacionada com a inteligência computacional (IC) ou aprendizagem de máquina. Se esse é de fato o objetivo das redes inteligentes, a rede elétrica deve ser projetada para suportar tal inteligência. Para tal, é necessário compreender os aspetos de aprendizagem de máquina que provavelmente serão implementados na rede inteligente (BUSH, 2013).

No entanto, o termo, inteligência não tem sido claramente definido. Portanto, isso ainda deixa em aberto a noção do que realmente é a inteligência de máquina, como pode ser medida e em que grau pode ser utilizada na rede de energia elétrica. Hernández-Orallo e Dowe (2010) relata uma forma utilizada para quantificar a inteligência da máquina de maneira nova e interessante, conhecida como teoria da informação algorítmica e tem a sua inspiração a partir da complexidade de Kolmogorov, que relaciona a noção de complexidade de informação pelo tamanho do menor programa que calcula a informação.

Diversos trabalhos têm sido publicados sobre a aplicação de técnicas de IC em Sistemas

Elétricos de Potência (SEP) desde o aparecimento do artigo publicado em (WOLLENBERG, 1986). Os sistemas baseados em conhecimento, como são os Sistemas Inteligentes (SI), têm sido amplamente utilizados para auxiliar o trabalho de profissionais da área dos SEP, principalmente aqueles que exercem funções que precisam de experiência para efetuarem tomadas de decisões, (FORD, 1985). Especificamente na área dos SEP, onde as decisões devem ser tomadas baseando-se em um conjunto muito grande de informações relevantes, cuja interpretação depende a garantia do funcionamento contínuo do fornecimento do serviço de energia elétrica. Assim, torna-se óbvia a importância de oferecer ao profissional da área dos SEP, ferramentas com uma arquitetura que a torne apta a gerenciar aquelas informações.

Várias pesquisas reportam avanços na aplicação dos SI no auxílio à operação de SEP. A importância desta alternativa aumenta à medida que estes sistemas crescem e se tornam mais complexos, o que dificulta ao operador ter controle absoluto e seguro de todas as áreas envolvidas, diminuindo consideravelmente sua capacidade de tomada de decisões rápidas e eficientes sem a assistência de um operador externo (VALIQUETTE; TORRES; MUKHEDKAR, 1991).

Considerando a complexidade dos SEP, e com o advento das subestações automatizadas, os medidores inteligentes, as *smart grids*, e a carência de ferramentas numéricas tradicionais que analisem as causas das interrupções não programadas, faltas na rede de transmissão, perdas técnicas e não técnicas de energia , entre outras problemáticas, faz-se necessária uma abordagem não tradicional, que permita ao engenheiro ou operador do sistema, entender melhor o SEP com que trabalha, levando a realizar um gerenciamento ótimo do SEP.

Segundo (LAMBERT-TORRES et al., 1997), na medida em que os sistemas se tornam mais complexos, mais imprescindível é o fato de que a decisão tomada esteja certa, e mais difícil se torna para o profissional executar ações sem ajuda externa. Dado o grande número de mudanças de estado operacional dos equipamentos, torna-se muito difícil para o engenheiro ou operador do sistema ter uma imagem de um SEP sem as informações em indicadores representativos. Portanto, faz-se necessário desenvolver SIs capazes de auxiliar na tomada de decisões, segundo um aprendizagem baseado no conhecimento de um especialista humano.

Considerando que tanto as concessionárias quanto as distribuidoras de energia elétrica utilizam sistemas que podem variar de uma concepção simples, até grandes aplicações mais complexas de coleta e análise de dados, e comando de grandes processos industriais, os dados obtidos (estado de equipamentos, valores de variáveis, alarmes, ultrapassagem de limites, identificador dos usuários, etc) são armazenados e, com o passar do tempo, formam grandes bancos de dados cheios de eventos e valores de variáveis ocorridos no processo sob supervisão e controle. Esses históricos de dados formam a base para análise do comportamento dos processos. O crescimento do volume de dados cria a necessidade de novas técnicas e ferramentas capazes de transformar, de forma eficiente, automática e inteligente, esses dados em informações significativas e em conhecimento. Essas informações, de grande importância para o planejamento,

controle, gestão e tomadas de decisão, podem estar implícitas ou escondidas sob um universo de dados, e não podem ser descobertas ou, facilmente identificadas utilizando somente sistemas convencionais de gerenciamento de banco de dados. Nesse contexto, é necessário a utilização de metodologias ágeis que permitam realizar a análise de grandes quantidades de dados de forma rápida e confiável, para assim lograr à automação de tarefas que precisam da inteligência e da velocidade que pode ser aplicada mediante o uso da IC.

O objetivo das pesquisas em IC é capacitar o computador a executar funções que são desempenhadas pelo ser humano usando conhecimento e raciocínio. Para que isto seja possível é necessário que sejam analisados todos os aspetos relativos ao desenvolvimento e uso da inteligência. Dentro desse contexto, a capacidade humana de agir inteligentemente é frequentemente associada ao conhecimento que se possui. Assim, torna-se evidente que a incorporação de conhecimento e raciocínio são os requisitos fundamentais para a construção de sistemas computacionais inteligentes.

1.1 CONTEXTO

A pergunta é, o que a IC pode fazer hoje? É difícil dar uma resposta concisa porque existem muitas atividades em vários sub-campos. Nesta seção são analisadas algumas aplicações da IC na área de engenharia elétrica, como pode ser visto a seguir:

Sistema SCADA: Um sistema especialista de Supervisão, Controle e Aquisição de Dados (SCADA), é um sistema utilizado para coletar, armazenar, analisar e apresentar os dados de uma subestação de energia elétrica, sob supervisão e controle. Em alguns desses sistemas, em função dos dados coletados e analisados, algumas decisões são tomadas e enviadas aos atuadores que interferem no processo controlado em tempo real. O sistema SCADA é um processo comum de aplicação e controle, que adquire dados do processo por meio de estações remotas e os enviam para processamento por um computador central. Uma aplicação do sistema especialista SCADA pode ser vista no trabalho apresentado por Oliveira, Braz e Ferreira (1994). Neste trabalho é desenvolvido um sistema especialista protótipo para tratamento de alarmes (SETA) para sistemas de tipo SCADA. Este sistema é baseado no conhecimento dos operadores do sistema elétrico. Um simulador de eventos foi desenvolvido para gerar ocorrências de alarmes, as quais podem ser de três tipos: aleatória, manual ou através de um arquivo de alarmes previamente montado. Através deste último modo de simulação do alarme, é possível reproduzir situações de emergência ocorridas num sistema elétrico real e os resultados obtidos com a utilização do sistema SETA foram bastantes satisfatórios.

Sistema Inteligente para tomada rápida de decisões nos sistemas elétricos: A ideia deste sistema inteligente desenvolvido por Filho (2006) foi auxiliar os operadores de uma concessionária de energia elétrica brasileira na recomposição sistêmica da rede. Para isto, foi desen-

volvida uma metodologia que utiliza, além de uma base de conhecimento própria, a integração com os sistemas da concessionária. Também foi utilizado um sistema extrator de conhecimento de grandes bases de dados que funciona em conjunto com um sistema especialista. Este sistema inteligente hierárquico aciona rotinas computacionais externas de apoio, bem como bases de dados existentes na concessionária.

Aplicações da técnica de mineração de dados para previsão da estabilidade transitória nos sistemas elétricos de potência: Tao et al. (2004) apresenta uma estrutura de mineração de dados para a simulação dos dados históricos das unidades de medidas e posteriormente realizar a previsão da estabilidade transitória nos sistemas de energia elétrica.

Mineração de dados para a detecção de barras sensíveis e barras influentes em um sistema de potência sujeito a distúrbios: Segundo Tso et al. (2004), muitos tipos de perturbações nos sistemas de potência podem conduzir a uma redução da carga do sistema. Neste trabalho uma técnica de Mineração de Dados (MD) foi aplicada a um SEP em Hong Kong para detetar as subestações mais sensíveis a distúrbios. Além disso, baseados nas análises de correlação de perfil de tensão, são identificadas as barras mais influentes onde o ajuste mais eficaz de tensão pode ser estrategicamente aplicado para auxiliar o barramento sensível, e a recuperação de tensão decorrente da perturbação pode ser deduzida.

Uma abordagem probabilística para o planejamento de rede nos sistemas elétricos de potência sob incertezas: Vassena et al. (2003) apresentam neste trabalho a utilização de uma técnica de MD para analisar uma base de dados com grande quantidade de cenários simulados usando Monte Carlo. O objetivo é utilizar a metodologia de MD para realizar os estudos de planejamento das redes de distribuição de energia elétrica sob incertezas a longo prazo. Nesta abordagem as principais incertezas externas durante o planejamento horizontal são modeladas como macro-cenários em diferentes instantes de tempo futuros. Técnicas de MD, são aplicadas para extrair informações do banco de dados, de modo a classificar cenários e reforços da rede de acordo com diferentes critérios.

Um Hardware de detecção de anomalias em tempo real utilizando uma Rede Neural Artificial para Proteção de Distância: Venkatesan e Balamurugan (2001) neste trabalho os pesquisadores descrevem um detector de falhas em tempo real para à aplicação de proteção a distância, baseado em redes neurais artificiais. Uma estrutura de rede neural ótima com um tempo de treinamento de curta duração é apresentada, para assim alcançar o objetivo em curto prazo.

Localização de falhas nas linhas de transmissão usando redes neurais de domínio complexo: Alves, Lima e Souza (2012) a localização da falha é uma tarefa crítica quando um distúrbio severo é causado por falha de isolamento em uma linha de transmissão. Para evitar custos económicos e sociais adicionais devido a interrupções de carga, o diagnóstico de falha deve ser concluído o mais rápido possível. Considerando que os SI têm sido bem sucedidos em lidar com

problemas de diagnóstico de falhas, este artigo propõe a aplicação de redes neurais de domínio complexo para mapear a relação entre sinais elétricos e localização de falha em linhas de transmissão. As redes neurais de domínio complexo permitem a representação de tensão/corrente sem amplitude e fase arbitrariamente desacopladas. Além disso, são analisados vários esquemas de representação de tensão e corrente, com base em informações eletromagnéticas transitórias e estacionárias. Para fins comparativos, essas representações de entrada também são testadas com redes neurais de domínio real. Os testes consideram condições realistas de operação/falha e assumem que a classificação de falhas já foi tratada.

Gerenciamento de energia de um veículo elétrico híbrido baseado em aprendizagem de máquina para minimizar o custo total operacional: Lin et al. (2015) neste artigo é estudado o problema de gerenciamento de energia em veículos elétricos híbridos (VEHs) com foco na minimização do custo operacional de um VEH, incluindo o custo de reposição de combustível e bateria. Mais precisamente, o artigo apresenta uma pesquisa na qual é utilizado o aprendizagem de forma aninhado em que tanto as ações ótimas (que incluem a seleção da relação de transmissão e o uso do motor de combustão interna versus o motor elétrico para conduzir o veículo) e os limites na faixa do estado de carga da bateria são aprendidas sobre a marcha. O ciclo interno do processo de aprendizagem é a chave para a minimização do uso de combustível, ao passo que o ciclo externo do processo de aprendizado é crítico para minimizar o custo de reposição da bateria amortizada. Os resultados experimentais demonstram uma redução de custos operacionais máxima de 48% pela política de gerenciamento de energia de VEH proposta neste trabalho.

Modelagem elétrica da dinâmica da bateria de um veículo elétrico usando máquinas do vetor de suporte: Majid et al. (2013) As baterias utilizadas em veículos elétricos (VE) são muitas vezes limitadas pela capacidade. Quanto tempo você pode usar um VE é determinado pela duração da bateria. No entanto, há dificuldade em modelar a relação entre a tensão de carga e a corrente sob diferentes temperaturas e estados de carga (State of Charge - SOC), por conta da não-linearidade das propriedades da bateria. O objetivo da pesquisa apresentada neste artigo é modelar a dinâmica de carregamento da bateria usando a máquina de vetor de suporte, que é uma poderosa ferramenta para se aproximar da função não linear. Uma bateria de 1000 mAh Li-Po/MH é utilizada para obter o modelo da bateria e criar a base de dados utilizados nos testes de aprendizagem da máquina de vetor de suporte. Observa-se que o modelo da máquina de vetor de de suporte pode simular a dinâmica de carregamento da bateria com quantidades de dados experimentais limitados, obtendo como resultado final um baixo erro relativo.

Além destas aplicações, existe uma grande quantidade de pesquisas que estão sendo desenvolvidas na área de IC aplicada à solução de problemas de engenharia elétrica, uma ampla revisão bibliografica sobre este tema pode ser encontrada em (BUSH, 2013).

Nesta tese são abordados quatro tipos de problemas diferentes da área da engenharia elétrica

e têm sido aplicados diferentes métodos de IC que têm permitido chegar a soluções aceitáveis quando comparadas com soluções calculadas mediante o uso de modelagem matemática.

1.2 OBJETIVOS

1.2.1 Objetivo geral

Desenvolver métodos derivados da IC baseados em Mineração de Dados e Sistemas Inteligentes, para resolver problemas relacionados com gerenciamento de energia no SDEE e no STEE, de forma a auxiliar ao operador do sistema na tomada de decisões.

1.2.2 Objetivos específicos

1. Utilizar a ferramentas computacional WEKA, que permitam o rápido treinamento dos SIs, para realizar uma aplicação dos mesmos de forma eficiente para a extração do conhecimento em bancos de dados, sendo uma solução rentável para as concessionárias.
2. Aplicar SIs para resolver o problema de controle centralizado Volt/VAr em modernos SDEE usando medições elétricas.
3. Usar a mineração de dados para a identificação de fraudes nas redes de distribuição de energia elétrica, analisando padrões de consumo;
4. Aplicar SIs para resolver o problema de localização de faltas nas redes de transmissão de energia elétrica, usando histórico de medições das faltas na rede.
5. Utilizar algoritmos de IC para realizar à coordenação de carga ótima de VEs e de dispositivos de armazenamento de energia em um SDEE.

1.3 CONTRIBUIÇÕES DA TESE

As principais contribuições deste trabalho são apresentadas a seguir:

- A aplicação de três SIs para resolver o problema de controle centralizado Volt/VAr em modernos SDEE usando medições elétricas.
- A aplicação de algoritmos de mineração de dados para resolver o problema de detecção de fraudes nas redes de distribuição de energia elétrica, analisando padrões de consumo.
- Aprimoramento das técnicas atualmente utilizadas pelas empresas concessionárias de energia elétrica na detecção de fraudes no SDEE.

- A aplicação de SIs para resolver o problema de localização de faltas nas redes de transmissão de energia elétrica, usando histórico de medições das faltas na rede.
- Um modelo flexível de RNA, MVS e AD para análise da otimização do carregamento de VEs e de dispositivos de armazenamento de energia ao longo do dia.
- A exploração científica da IC e da MD como ferramentas para descoberta de conhecimento no domínio de distribuição e transmissão de energia elétrica, permitindo com que a tomada de decisões possa ser feita em tempo real.

1.4 ESTRUTURA DA TESE

Esta tese de doutorado está organizada em sete capítulos, distribuídos da seguinte forma:

- No Capítulo 2 é apresentada a revisão bibliográfica sobre IC e suas aplicações na solução de problemas na área da Engenharia Elétrica.
- No Capítulo 3 é desenvolvida a pesquisa para aplicação de SIs ao problema de controle centralizado Volt-VAr em modernos SDEEs.
- No Capítulo 4 é apresentada a aplicação de MD como solução ao problema de detecção de fraudes nas redes de distribuição de energia elétrica.
- No Capítulo 5 é apresentada a aplicação de algoritmos de IC como solução ao problema de localização de faltas em STEE.
- No Capítulo 6 é desenvolvida a pesquisa para aplicação de algoritmos de IC para realizar à coordenação de carga ótima de VEs e de dispositivos de armazenamento de energia em um SDEE.
- No Capítulo 7 apresentam-se as conclusões da tese e as propostas para trabalhos futuros.
- No apêndice A apresenta-se o software de livre distribuição WEKA, que foi utilizado para realizar os testes.

2 INTELIGÊNCIA COMPUTACIONAL

Este capítulo considera a revisão bibliográfica da área relacionados com IC, tema ao qual está relacionado esta tese. Para uma melhor compreensão do estado da arte, o capítulo consta de quatro seções. Na Seção 2.1, é apresentada a introdução à IC baseada na literatura especializada. Na Seção 2.2, são apresentadas as principais características dos sistemas inteligentes (SI). Na Seção 2.3, são apresentados os paradigmas de aprendizagem em IC. Finalmente, na Seção 2.4, são apresentados os passos a serem realizados durante o processo de mineração de dados (MD).

2.1 INTRODUÇÃO

A IC é um dos campos mais recentemente abordados em ciências e engenharia. Na atualidade, a IC abrange uma enorme variedade de sub-campos, dentro dos quais podem ser encontrados, aprendizagem, percepção, e tarefas específicas, tais como demonstração de teoremas matemáticos, jogos de xadrez, diagnóstico de doenças, planejamento autônomo e escalonamento, planejamento logístico, robótica, tomada de decisões diante de incertezas, resolução de problemas complexos (de difícil formulação matemática envolvendo muitas variáveis), jogos de tabuleiro, direção de carros, entre outros.

A IC ou também conhecida na literatura especializada como inteligência artificial, é uma área da ciência relevante para qualquer tarefa intelectual, é por isto que pesquisadores desta área afirmam que a IC é um campo universal (RUSSELL; NORVIG, 2009). Na literatura podem ser encontradas varias definições de IC, algumas delas são:

Segundo Nilsson (1998) *"A IC está relacionada a um desempenho inteligente de artefactos"*; Winston (1992) afirma que *"a IC é o estudo das computações que tornam possível perceber, raciocinar e agir"*; outra definição foi dada por Rich e Knight (1991) *"A IC é o estudo de como os computadores podem fazer tarefas que hoje são melhor desempenhadas pelas pessoas"*, por outro lado, o Bellman (1978) define a IC como *"Automatização de atividade que associamos ao pensamento humano, atividades como a tomada de decisões, a resolução de problemas, o aprendizado..."*; também são encontradas definições como a de Kurzweil (1990) quem define a IC como *"a arte de criar máquinas e dispositivos de controle que executam funções que exigem inteligência quando executadas por pessoas"*. Todas estas e outras definições têm sido seguidas para o estudo da IC, cada uma delas por pesquisadores diferentes com métodos diferentes.

Segundo as diversas definições que têm sido dadas por diferentes pesquisadores, o objetivo

científico da IC é o entendimento dos princípios que induzem o comportamento inteligente em sistemas naturais e artificiais. A hipótese principal é que o raciocínio equivale à computação. Segundo esta hipótese o objetivo mais prático é a especificação de métodos para projetos de artefactos inteligentes e úteis, onde o propósito consiste em entender como o comportamento inteligente é possível; cabe aclarar que, o objetivo não é simular comportamentos inteligentes, o objetivo é entender SI (naturais ou sintéticos) por meio de sintetização destes sistemas. A metodologia está no projeto, construção e experimentação de sistemas computacionais capazes de executar tarefas tipicamente vistas como inteligentes.

Ao longo dos últimos anos de história da ciência, a ênfase tem sido voltada nos algoritmos como o assunto principal de estudo. No entanto, alguns pesquisadores da área de IC sugerem que, para muitos problemas, faz mais sentido preocupar-se com os dados e ser menos exigentes sobre qual algoritmo aplicar (RUSSELL; NORVIG, 2009).

Isto é verdade devido à disponibilidade crescente de fontes de dados muito grandes: por exemplo, grandes quantidades de medidas de consumo de energia elétrica fornecidos por medidores inteligentes, medições de dispositivos instalados ao longo do SEP, entre outras cifras que crescem diariamente e fazem com que o manuseio desta informação seja cada vez mais difícil, já que, torna-se inviável para qualquer ramo de negócio ou pesquisa, investigar grandes volumes de informação utilizando só pessoas, mesmo que, a equipe de trabalho disponível seja muito grande.

Segundo Goldschmidt, Passos e Bezerra (2015), dados científicos usados em projetos de pesquisa, têm alcançado proporções gigantescas. Estimasse uma taxa de crescimento de dados mundial em torno de 40% ao ano na próxima década, o que, em 2020 devera alcançar um total de cerca de 44 zettabytes de informações digitais em todo o mundo. Este crescimento tem sido significativo graças ao advento da Internet Das Coisas (IDC) e das redes inteligentes.

O principal aporte da IC veio a ser corroborado com a necessidade de extrair conhecimento e processar dados de forma rápida e confiável, onde seu manuseio seria humanamente impraticável. Goldschmidt, Passos e Bezerra (2015) afirma que o valor dos dados armazenados está diretamente ligado à capacidade de se extrair conhecimento de alto nível a partir deles, ou seja, informação útil que sirva de apoio à tomada de decisões ou para exploração e melhor entendimento da ocorrência geradora dos dados.

É um pouco irónico dizer que é necessário usar a classificação e algoritmos de Aprendizado de Máquina (AM) para dar sentido a eles, porque é assim que os algoritmos de AM funcionam normalmente. Há três características gerais de algoritmos de AM: representação, avaliação e otimização (BUSH, 2013).

- **Representação:** é como um classificador é representado em um formato legível por máquinas. Os exemplos incluem casos simples, hiperplanos, árvores de decisão, redes neu-

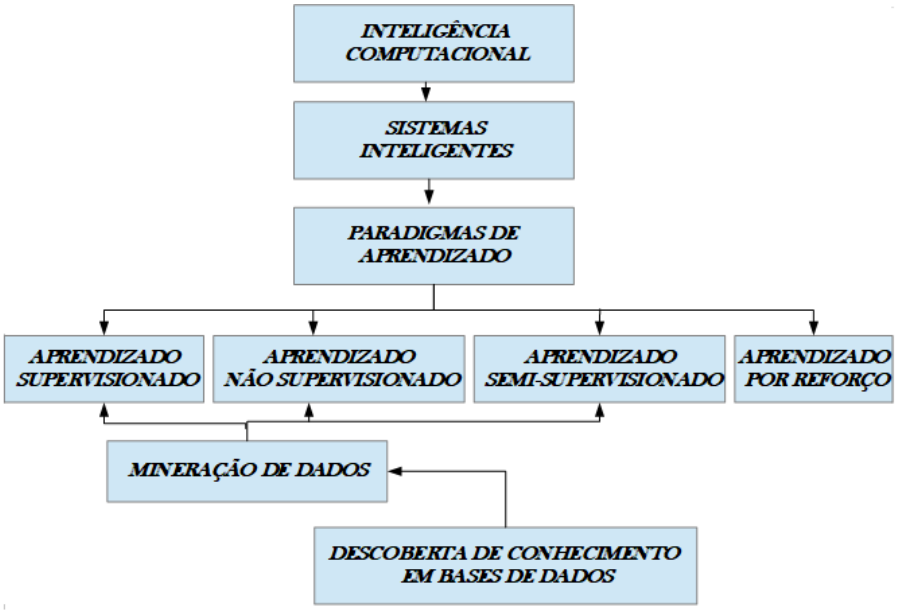
ronais e modelos gráficos. A escolha da representação coloca limites sobre o que pode e o que não pode ser aprendido, ou simplesmente, se pode ou não pode ser representado (BUSH, 2013).

- **Avaliação:** refere-se à escolha de uma função objetivo que é utilizada para distinguir o desempenho dos classificadores. Exemplos incluem a exatidão / taxa de erro, precisão, erro quadrático, ganho de informação e divergência (BUSH, 2013).
- **Otimização:** é o algoritmo usado para procurar entre os classificadores aqueles com melhor desempenho. Exemplos incluem otimização combinatória, tais como *branch-and-bound* e otimização contínua (BUSH, 2013).

Considerando que as técnicas de IC permitem encontrar e interpretar padrões em grandes bases de dados implementando habilidades do ser humano em sistemas computacionais inteligentes, cabe destacar que, os SIs têm um papel muito importante por estarem aptos a lidar com grandes bases de dados e serem aplicados em diferentes tipos de pesquisas.

A Figura 1, ilustra a taxonomia das metodologias utilizadas dentro da área de IC para resolver diversos problemas relacionados com a automação de processos de aprendizagem.

Figura 1 - Taxonomia das metodologias utilizadas dentro da área de IC.



Fonte: Próprio autor

Na Figura 1 pode ser observado que os SIs ocupam um lugar muito importante dentro desta hierarquia, portanto na seguinte seção vai ser abordado este tema com mais profundidade.

2.2 SISTEMAS INTELIGENTES

Dentro do contexto dos SIs é necessário falar do que é o "conhecimento". Portanto, algumas hipóteses e delimitações são necessárias. Em primeira instância, pode-se pensar em níveis de conhecimento: fatos, conceitos, regras e metarregras. O conhecimento pode ser representado como uma combinação de estruturas de dados e procedimentos interpretativos que levam a um comportamento conhecido, o qual fornece informações a um sistema que pode planejar e decidir (GOMES et al., 2014).

O conhecimento está dividido em vários tipos:

- Conhecimento procedimental;
- Conhecimento declarativo;
- Conhecimento de senso comum;
- Conhecimento heurístico.

O tipo de conhecimento necessário à solução dos problemas existentes determina quais fontes de informação serão utilizadas pelos indivíduos. Portanto, o conhecimento pode ser gerado a partir da combinação de diferentes informações. Assim, uma decisão pode ser tomada por meio de análise lógica ou pode estar apoiada em dados heurísticos ou intuitivos (RAILEANU; STOFFEL, 2004).

Os pontos chaves nos SIs são:

1. Habilidade para usar conhecimento visando desempenhar tarefas ou resolver problemas.
2. Capacidade para aproveitar associações e inferências para trabalhar com problemas complexos que se assemelham a problemas reais.

Uma característica relevante dos SIs é que têm a habilidade de armazenar e recuperar eficientemente grandes quantidades de informação, para resolver problemas ou tomar decisões. O comportamento inteligente de um sistema é resultado de múltiplas e encadeadas decisões. A escolha da melhor decisão tomada, ou controle de decisão, é baseada em critérios de desempenho, duração e risco. (CHEN-CHING; PIERCE; SONG, 1997)

Cabe salientar, que os SIs podem ser desenvolvidos usando diversas técnicas, as quais podem ser aplicadas isoladamente ou em conjunto para auxiliar o processo decisório. As principais técnicas e metodologias usadas pelos SIs são: Aquisição de Conhecimento, AM, RNAs, Lógica Fuzzy, MVSs, Computação Evolutiva, Agentes, ADs, MD e mineração de textos. Cada

uma dessas técnicas oferece uma variedade de graus de habilidade para representar o conhecimento humano. Para o desenvolvimento desta pesquisa foram estudadas técnicas de MD e o AM fazendo uso de algoritmos de RNAs, ADs e MVSs.

A construção de SIs capazes de aprender por experiência, tem sido objeto de discussões tanto técnicas quanto filosóficas. Porém, relevantes pesquisas na área têm demonstrado que máquinas podem apresentar um significante nível de aprendizagem, mesmo não estando claramente definidas as fronteiras dessa habilidade de aprendizagem. Estas habilidades dependem dos paradigmas de aprendizagem de máquina que são utilizados, os quais serão descritos na seguinte secção.

2.3 PARADIGMAS DE APRENDIZAGEM

Técnicas de IC, em particular de AM, têm sido utilizadas com sucesso na solução de um grande número de problemas reais, incluindo problemas da área de SEP. Segundo Russell e Norvig (2009), para alguns sistemas de aprendizagem é necessário predizer se uma certa ação irá fornecer uma certa saída, para isto, é necessário conhecer as diferenças entre os vários tipos de aprendizagem.

Os algoritmos utilizados na área de IC envolvem aprendizagem de exemplos com o objetivo de generalização. Isso significa que a solução deve aprender a reconhecer as características gerais do que é "pensado"ser importante e não detalhes específicos do conjunto de treinamento. Não é suficiente aprender um conjunto de treinamento perfeitamente, porque o objetivo é aprender a classificar corretamente informações que nunca antes foram "vistas" pelo classificador (BUSH, 2013).

A aprendizagem de uma função geral ou regra a partir de pares específicos de entrada-saída é chamada de aprendizagem indutiva; em quanto que, passar de uma regra geral conhecida a uma nova regra derivada, porém útil, por permitir um processamento mais eficiente é chamada de aprendizagem analítica ou dedutiva. Nesse sentido, é possível classificar os principais tipos de aprendizagem de máquina da seguinte forma:

- Aprendizagem supervisionada;
- Aprendizagem não supervisionada;
- Aprendizagem semissupervisionada;
- Aprendizagem por reforço.

Aprendizagem supervisionada: na qual dado um conjunto de observações ou exemplos rotulados, isto é, conjunto de observações em que a classe de cada exemplo é conhecida, o

objetivo é encontrar uma hipótese capaz de classificar novas observações entre as classes já existentes; em outras palavras, o agente inteligênte observa alguns exemplos de pares de entrada e saída, e aprende uma função que faz o mapeamento de entrada para a saída Bishop (2007).

Aprendizagem não supervisionada: a qual dado um conjunto de observações ou exemplos não rotulados, o objetivo é tentar estabelecer a existência de grupos ou similaridades nesses exemplos. Neste tipo de aprendizagem são aprendidos padrões na entrada, embora não seja fornecido nenhum feedback explícito. A tarefa mais comum de aprendizagem não supervisionada é o agrupamento, isto é, a deteção de grupos de exemplos de entrada potencialmente úteis Bishop (2007).

Aprendizagem semissupervisionada: a qual dado um pequeno conjunto de observações ou exemplos rotulados e não rotulados, o objetivo é utilizar ambos os conjuntos para encontrar uma hipótese capaz de classificar novas observações entre as classes já existentes. A aprendizagem semissupervisionada é um meio termo entre aprendizagem supervisionada e não supervisionada Bishop (2007).

Aprendizagem por reforço: a qual o agente aprendiz interage com o meio ambiente que o cerca e aprende uma política ótima de ação por experimentação direta com o meio. Dependendo de suas ações, o agente aprendiz é recompensado ou penalizado. O objetivo do aprendiz é desenvolver uma política ótima que maximize a quantidade de recompensas recebidas ao longo de sua execução Bishop (2007).

Cabe salientar que, para cada tipo de aprendizagem existem diferentes tipos de paradigmas, dentre os quais são destacados:

- O paradigma simbólico;
- O paradigma conexionista;
- O paradigma estatístico;
- O paradigma genético;
- O paradigma baseado em protótipos.

Durante o desenvolvimento desta pesquisa foram utilizados os paradigmas simbólico, conexionista e estatístico, por isto, estes três paradigmas serão explicados mais profundamente nas seguintes seções.

2.3.1 Paradigma simbólico

Os sistemas de aprendizagem simbólica buscam aprender construindo representações simbólicas de um conceito por meio da análise de exemplos e contra-exemplos desse conceito. As

representações simbólicas estão tipicamente representadas na forma de alguma expressão lógica, árvores de decisão (AD), regras de produção ou redes semânticas. Entre as representações simbólicas mais estudadas e mais pesquisadas na literatura estão as AD.

2.3.1.1 Aprendizagem em AD

Uma AD é uma estrutura de dados definida recursivamente como um nó folha que corresponde a uma classe e *m* nós de decisão que contém um teste sobre algum atributo. Para cada resultado do teste existe uma aresta para uma sub-árvore. Cada sub-árvore tem a mesma estrutura que a árvore. As ADs são utilizadas como modelo preditivo, (MAHMOOD; SATULURI; KUPPA, 2011), podendo realizar classificação e representação de modelos de regressão. As ADs são categorizadas como um método de treinamento supervisionado para encontrar uma conexão lógica entre atributos de entrada e o propósito dos atributos que representam a conexão lógica em estruturas como um modelo, Rokach e Maimon (2008).

As ADs são uma das técnicas mais utilizadas, sendo muito popular entre as pesquisas pela sua simplicidade, inteligibilidade e facilidade de desenvolvimento. A construção de ADs é mais rápida e tem maior precisão que outros algoritmos de classificação.

É fácil perceber que a árvore pode ser representada como um conjunto de regras. Cada regra tem seu inicio na raiz da árvore e caminha até uma de suas folhas. A chave para o sucesso de um algoritmo de aprendizado por ADs depende do critério utilizado para escolher o atributo que divide o conjunto de exemplos em cada iteração Russell e Norvig (2009).

Uma AD representa uma função que toma como entrada um vetor de valores de atributos e retorna uma decisão, isto é, um valor de saída único. Os valores de entrada e saída podem ser discretos ou contínuos. Uma AD alcança sua decisão executando uma sequência de testes. Cada nó interno na árvore corresponde a um teste do valor de um dos atributos de entrada. Cada nó de folha na árvore especifica o valor a ser retornado pela função.

Para uma grande variedade de problemas, o formato da AD gera um resultado conciso. Contudo, algumas funções não podem ser representadas de forma concisa, portanto, as ADs são boas para alguns tipos de funções e ruins para outros. O algoritmo de aprendizagem em AD adota uma estratégia gulosa de dividir para conquistar: sempre testar o atributo mais importante em primeiro lugar. Esse teste divide o problema em subproblemas menores que podem então ser resolvidos de forma recursiva. Entenda-se como atributo mais importante aquele que faz maior diferença para a classificação de um exemplo.

A metodologia de ADs pode ser aplicada através de diversos algoritmos, entre os quais estão: J48, REPTree, PART, Ridor, JRip, ID3, ASSISTANT, CART, entre outros. Nesta pesquisa, foi utilizado o algoritmo J48 que consegue uma probabilidade de acerto superior aos demais classificadores, assim como apresentado na pesquisa realizada por Witten e Frank (2011).

O algoritmo J48 é uma versão do tradicional algoritmo C4.5, desenvolvido por Quinlan (1996). O algoritmo C4.5 consegue, através da discretização de valores numéricos através de heurísticas, processar atributos de qualquer tipo de entrada. No entanto, a árvore não é capaz de realizar regressão e desempenha somente o papel de classificador, isto é, o atributo de classe deve ser discreto Zhao et al. (2007).

Segundo o Zhao et al. (2007), o algoritmo C4.5 realiza a classificação de um conjunto de dados. Mediante uma divisão recursiva dos dados, as ADs crescem usando uma estratégia conhecida como *Depth-first*. Com o intuito de reunir o ganho de entropia de todos esses testes binários de forma eficiente, o conjunto de dados de treinamento pertencentes a cada nó da árvore é ordenado para os valores dos atributos contínuos, e os ganhos de entropia do corte binário com base em cada valor distinto, estão calculados em uma varredura dos dados ordenados. Este processo é repetido para cada atributo contínuo.

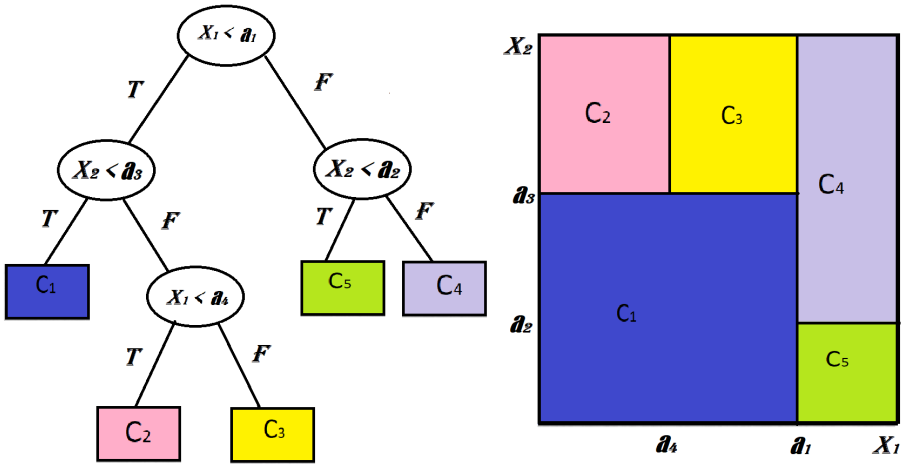
Em um algoritmo de aprendizagem em AD, o conjunto de exemplos é muito importante para a construção da árvore, mas, graficamente na árvore os exemplos não aparecem em nenhum lugar. Uma árvore é composta de testes em atributos no interior dos seus nós, valores de atributos nas ramificações e valores de saída nos nó folha (RUSSELL; NORVIG, 2009).

Uma AD é um grafo acíclico direcionado em que cada nó é um nó de divisão com dois ou mais sucessores, ou um nó folha.

- **Nó folha:** Também conhecida como nó da *função*. São considerados os valores da variável objetivo nos exemplos que chegam a um nó folha. Para casos simples, a função é a constante que minimiza a função de custo. Para problemas de classificação, essa constante é a *moda*. Para problemas de regressão, a constante que minimiza a função de custo do erro médio quadrático é a *média*, enquanto para a função de custo de desvio absoluto é a *mediana* (GAMA et al., 2011).
- **Nó de divisão:** Este nó contém um teste condicional baseado nos valores do atributo, onde as condições envolvem um único atributo e valores no domínio desse atributo (GAMA et al., 2011).

A Figura 2, ilustra uma AD e o espaço correspondente aos atributos X_1 e X_2. Cada nó da AD corresponde a uma área dentro do espaço em que estão alocados os objetos. A soma de todas as folhas (regiões) equivale a área total do espaço ocupado pelos objetos. Cabe salientar que, as áreas definidas pelas folhas da AD são mutuamente excludentes.

Figura 2 - AD e as regiões de decisão no espaço de objetos.

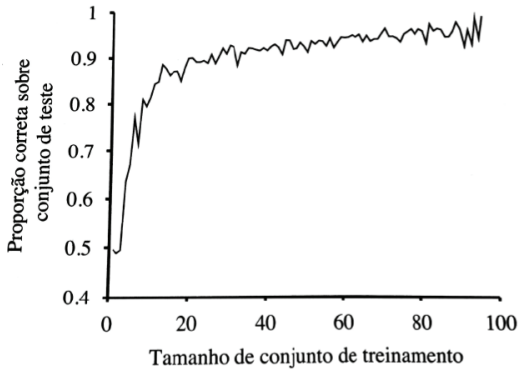


Fonte: Adaptado de Gama et al. (2011)

Existe o problema de super-interpretar a árvore que o algoritmo seleciona. Quando há diversas variáveis de importância similar, a escolha entre elas é um tanto arbitrária: com exemplos de entrada ligeiramente diferentes, em primeiro lugar, uma variável diferente para dividir, e a árvore toda pareceria completamente diferente. Em quanto a função calculada pela árvore ainda seria semelhante, mas a estrutura da árvore poderia variar muito.

Segundo Russell e Norvig (2009) a precisão de um algoritmo de aprendizagem pode ser avaliada com uma curva de aprendizagem, como mostrado na Figura 3, a curva mostra que, à medida que o tamanho do conjunto de treinamento cresce, a precisão aumenta.

Figura 3 - Curva de aprendizagem do algoritmo de AD em 100 exemplos gerados aleatoriamente.



Fonte: Adaptado de Russell e Norvig (2009)

2.3.1.2 Escolha de atributos

A busca gulosa utilizada em aprendizagem de AD foi projetada para minimizar aproximadamente a profundidade da árvore final. Buscando assim, escolher o atributo que vá o mais longe possível na tentativa de fornecer uma classificação exata dos exemplos. Em geral, depois que o primeiro teste de atributo separa os exemplos, cada resultado será um novo problema de aprendizagem de AD em si, com menos exemplos e com um atributo a menos. Um atributo perfeito divide os exemplos em conjuntos, cada um dos quais pode ser todo positivo ou negativo, e posteriormente passar a se tornar as folhas da árvore.

Para isto acontecer é necessário utilizar algum tipo de medida formal, e pode ser usada a noção de ganho de informação, que é definida em termos de entropia Shannon e Weaver (1971), isto é, a quantidade fundamental em teoria da informação. A entropia é uma medida da incerteza de uma variável aleatória; a aquisição de informação corresponde a uma redução na entropia. No caso de uma variável aleatória com um único valor, (por exemplo uma moeda viciada que sempre dá cara) não há incerteza e, portanto, sua antropia é definida como zero; assim, obtém-se a informação observando seu valor.

2.3.1.3 Generalização e super-adaptação

Para alguns tipos de problemas, o algoritmo de aprendizagem em AD vai gerar uma grande árvore quando realmente não houver padrão a ser encontrado. Neste caso pode acontecer um fenômeno conhecido como super-adaptação. Isto ocorre com todos os tipos de algoritmos de aprendizagem, mesmo quando a função de destino não for aleatória. Segundo Russell e Norvig (2009) a super adaptação torna-se mais provável à medida que o espaço de hipótese e o número de atributos de entrada cresce, e menos provável à medida que é aumentado o número de exemplos de treinamento.

Para combater o problema de super-adaptação em algoritmos de aprendizagem em AD, é utilizada uma técnica chamada de *poda de AD* (MITCHELL, 1997). A poda é realizada através da eliminação de nós que não são claramente relevantes. Se inicia com uma árvore cheia, como a gerada pela aprendizagem em AD. Seguidamente, verifica-se um nó de teste que tem somente nós folhas como descendentes. Se o teste parece ser irrelevante, isto é, é detectado apenas o ruído dos dados, então o teste é eliminado, substituindo-o por um nó folha. Este processo é repetido, considerando cada teste apenas com descendentes folha, até que cada um seja podado ou aceito.

Agora a pergunta é, como é possível saber se um nó está testando um atributo irrelevante? Um bom indicativo para a irrelevância é considerar o ganho de informação. O tamanho do ganho que se deve exigir para fazer uma divisão baseada em atributo específico, é obtido usando um teste de significância estatístico. Este teste inicia pela suposição de que não existe ne-

nhum padrão subjacente (hipótese nula). Então, os dados reais são analisados para calcular até que ponto eles divergem de uma ausência perfeita de padrão. Se o grau de desvio for estatisticamente improvável, ele será considerado como boa evidência da presença de um padrão significativo nos dados. As probabilidades são calculadas a partir de distribuições-padrão da proporção de desvio que se esperaria ver em uma amostragem aleatória.

Com a poda, os ruídos nos exemplos podem ser tolerados. Os erros nos rótulos dos exemplos fornecem um aumento linear no erro de previsão. As árvores podadas apresentam um desempenho significativamente melhor do que as árvores não podadas quando os dados contêm grandes quantidades de ruídos. Além disso, as árvores podadas geralmente são muito menores e, portanto, mais fáceis de entender (RUSSELL; NORVIG, 2009).

2.3.1.4 Aplicabilidade de AD

Um sistema de aprendizagem em AD para aplicações reais deve ser capaz de manipular problemas tais como:

- **Omissão de dados:** Em muitos domínios, nem todos os valores de atributos são conhecidos para todo exemplo. Os valores podem não ter sido registrados ou talvez seja muito dispendioso obtê-los.
- **Atributos multi-valorados:** Quando um atributo tem muitos valores possíveis, a medida de ganho de informação fornece uma indicação imprópria da utilidade do atributo.
- **Atributos de entrada com valores contínuos e inteiros:** Os atributos de valores contínuos ou inteiros, tem um conjunto infinito de valores possíveis. Em vez de gerar um número infinito de ramificações, os algoritmos de aprendizagem de AD em geral encontram o ponto de divisão que fornece o mais alto ganho de informações.
- **Atributos de saída com valores contínuos:** Quando se tenta prever um valor numérico de saída, é possível usar as árvores de regressão, em vez de usar só uma árvore de classificação. Uma árvore de regressão tem em cada folha uma função linear de um subconjunto de atributos numéricos, em vez de um único valor.

Um sistema de aprendizagem em AD para aplicações reais deve ser capaz de manipular todos os problemas já mencionados. O tratamento de variáveis de valores contínuos é especialmente importante porque tanto processos físicos quanto financeiros fornecem dados numéricos. Em muitas áreas de pesquisa, da indústria e de comércio, as ADs costumam ser o primeiro método experimentado quando um método de classificação tem de ser extraído de um conjunto de dados. Uma propriedade importante das ADs é que possibilitam ao ser humano entender a razão de saída do algoritmo de aprendizagem. Essa propriedade não é compartilhada por algumas outras representações, tais como as redes neurais.

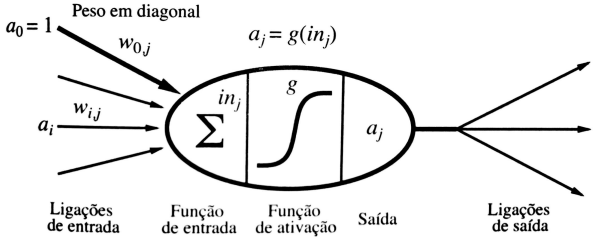
2.3.2 Paradigma conexionista

É uma das grandes linhas de pesquisa da IC e tem por objetivo investigar a possibilidade de simulação de comportamentos inteligentes através de modelos baseados na estrutura e funcionamento do cérebro humano. Com o renascimento do interesse sobre o estudo do conexionismo na década de 80, uma parte da pesquisa passa a se dedicar ao estudo de redes neurais, as quais vão ser descritas a seguir.

2.3.2.1 Redes Neurais Artificiais

Inspirados na hipótese de que a atividade mental consiste basicamente na atividade eletroquímica em redes de células cerebrais chamadas **neurônios**. Alguns dos trabalhos mais antigos e relevantes da IC tiveram o objetivo de criar RNAs baseados na interconexão de neurônios. Na Figura 4 pode ser visto o modelo matemático simples do neurônio desenvolvido por McCulloch e Pitts (1943).

Figura 4 - Modelo matemático simples de um neurônio.



Fonte: Adaptado de Russell e Norvig (2009)

Lembrando que uma RNA é apenas uma coleção de unidades conectadas, as propriedades da rede são determinadas pela sua topologia e pelas propriedades dos "neurônios". Desde seu início no ano de 1943 até a atualidade, têm sido desenvolvidos modelos muito mais detalhados e realistas, tanto de neurônios como de sistemas maiores no cérebro, levando ao campo moderno conhecido como neurociência computacional. Cabe salientar, que os pesquisadores de IC e os estatísticos tornaram-se interessados nas propriedades mais abstratas das RNAs, tais como sua capacidade de realizar computação distribuída, de tolerar entradas ruidosas e aprender. Embora outros tipos de sistemas tenham essas propriedades, as RNAs continuam sendo uma das formas mais populares e eficazes de aprendizagem.

As características que tornam a metodologia de RNA interessante do ponto de vista da solução de problemas são as seguintes:

- **Capacidade de aprender,** através de exemplos e de generalizar esta aprendizagem de

maneira a reconhecer instâncias similares que nunca haviam sido apresentadas como exemplo.

- **Bom desempenho em tarefas mal definidas,** onde falta o conhecimento explícito sobre como encontrar uma solução.
- **Elevada imunidade ao ruído,** isto é, o desempenho de uma RNA não entra em colapso nos casos em que se têm informações falsas ou ausentes, como é o caso dos programas convencionais, mas pode piorar de maneira gradativa.
- **Possibilidade de simulação de raciocínio a priori** e impreciso, através da associação com a lógica nebulosa.

Segundo Haykin (2008), as RNAs são modelos matemáticos que se assemelham às estruturas neurais biológicas e que têm capacidades adquiridas por meio de aprendizado e generalização. O aprendizado em RNAs consiste na fase em que a rede neural usa pares de dados de entrada e saída para modificar seus parâmetros de aprendizagem. Esta etapa pode ser considerada como uma adaptação da RNA às características intrínsecas de um problema, onde se procura cobrir um grande espectro de valores associados às variáveis pertinentes. Isto é feito para que a RNA adquira, através de uma melhora gradativa, uma boa capacidade de resposta para o maior número de situações possíveis. Além disso, a generalização de uma RNA está associada à sua capacidade de dar respostas coerentes para dados não apresentados a ela durante o treinamento. Portanto, espera-se que una RNA treinada tenha uma boa capacidade de generalização independentemente de ter sido controlada durante o treinamento (RUSSELL; NORVIG, 2009).

Existem dois aspetos básicos que caracterizam as RNAs: a arquitetura e o aprendizado. A arquitetura ou estrutura de uma RNA está relacionada com a forma como os neurônios estão conectados e com o tipo e quantidade de unidades de processamento. O aprendizado diz respeito às regras utilizadas para realizar o ajuste dos pesos da rede e da informação utilizada pelas regras (GAMA et al., 2011).

2.3.2.2 Estrutura das RNAs

As RNAs são compostas por nós ou unidades conectadas por ligações direcionadas. Uma ligação da unidade i para a unidade j serve para propagar a **ativação** a_i de i para j. Cada ligação tem um peso numérico $w_{i,j}$ associado a ele, que determina a força e o sinal de conexão. Cada unidade tem uma entrada fictícia $a_0 = 1$ com peso associado $w_{0,j}$. Cada unidade j primeiro calcula uma soma ponderada de suas entradas. Em seguida, aplica uma função de ativação g a essa soma para obter a saída.

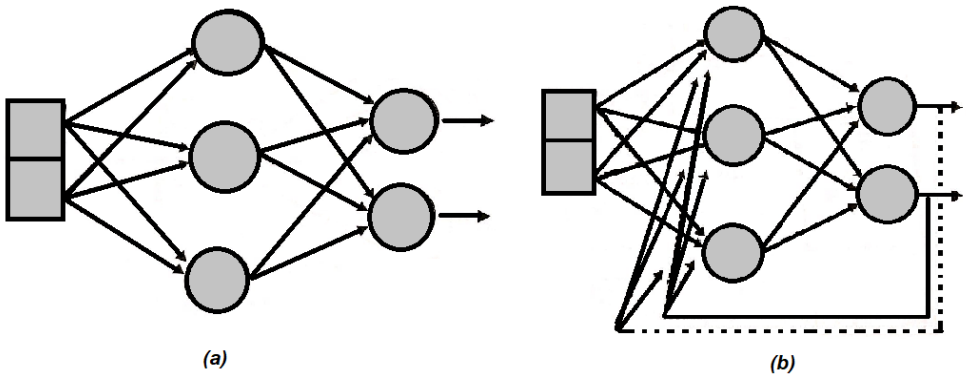
O funcionamento do modelo de uma RNA pode ser descrito intuitivamente da seguinte maneira: se a soma ponderada dos sinais de entrada de um neurônio ultrapassar um determinado limite de disparo, então a saída toma valor um (1). Se não ultrapassar, toma valor zero (0). A arquitetura da RNA é definida pela forma em que esses neurônios estão organizados e interconectados, ou seja, o número de camadas, o número de neurônios por camada, tipos de conexão entre os neurônios e a topologia da rede, como mostrado em Haykin (2008).

Uma vez definido o que é um neurônio, é possível estudar as propriedades de redes de neurônios interconectados, as chamadas RNAs. Existem duas formas fundamentalmente distintas para interconectar os neurônios. Uma é conhecida como rede com alimentação para frente mais comumente denominadas RNAs *feedforward*, tem conexões somente em uma direção, isto é, forma um grafo acíclico dirigido. Cada nó recebe a entrada de nós para cima e libera a saída de nós para baixo; não há laços. Uma rede com alimentação para frente representa uma função de sua entrada atual; portanto, não tem estado interno que não seja os próprios pesos. As redes com alimentação para frente estão dispostas em camadas, de tal forma que cada unidade recebe a entrada somente a partir de unidades na camada imediatamente anterior.

Por outro lado, existem as redes recorrentes. Estas redes alimentam suas saídas de volta às suas próprias entradas. Isto é, os níveis de ativação da rede formam um sistema dinâmico que pode atingir um estado estável ou apresentar oscilações ou um comportamento caótico. A resposta da rede para determinada entrada depende do seu estado inicial, que pode depender de entradas anteriores. Portanto, as redes recorrentes podem suportar memoria de curto prazo. Isso faz com que sejam mais interessantes como modelos de cérebro, mas também mais difícil de entender. Um caso limite de rede recorrente é a rede totalmente conectada onde cada neurônio está conectado a todos os outros e da qual toda estrutura de interligação é um caso particular.

A Figura 5 ilustra exemplos dos dois tipos de arquiteturas, uma recorrente e uma fedforward.

Figura 5 - (a) Rede neural *feedforward.* (b) Rede neural recorrente.



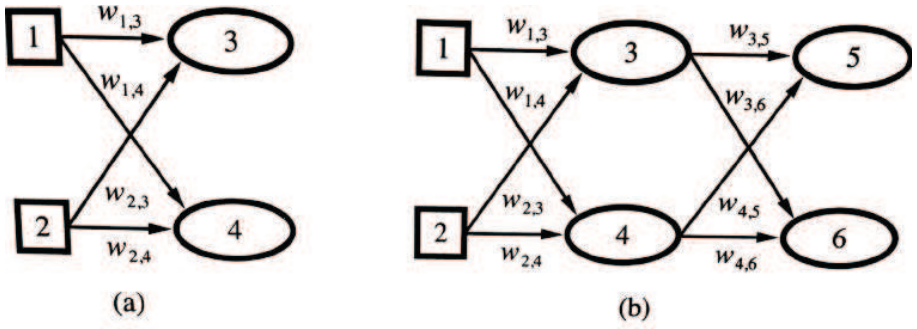
Fonte: Adaptado de Gama et al. (2011)

O número de camadas, o número de neurônios em cada camada, o grau de conectividade e a existência ou não de conexões de retropropagação definem a topologia de uma RNA (GAMA et al., 2011). Cabe salientar que nas últimas décadas foram desenvolvidas diversas arquiteturas de RNAs e algoritmos de treinamento. A seguir são apresentadas algumas arquiteturas de RNAs.

2.3.2.3 Rede Perceptron

Uma rede com todas as entradas conectadas diretamente com as saídas é chamada de rede neural de camada única ou rede perceptron. O primeiro modelo de rede neural -chamado *perceptron*- foi proposto por Rosenblatt (1962). Este modelo consiste em uma rede de duas camadas, uma de entrada e outra de saída, formadas por neurônios binários. Na Figura 6, a primeira coisa a notar é que uma rede *perceptron* com *m* saídas é realmente *m* redes separadas, pois cada peso afeta apenas uma saída. Assim, haverá *m* processos de treinamento separados.

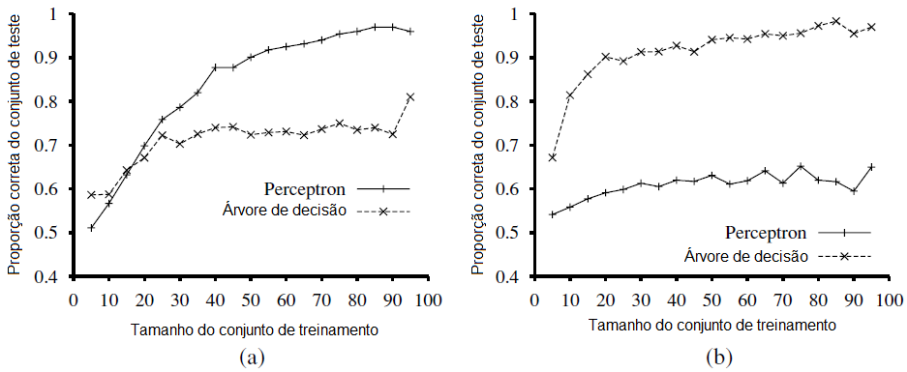
Figura 6 - (a) Rede *perceptron* com dois nós de entrada e dois nós de saída. (b) Rede neural com dois nós entradas, uma camada oculta de dois nós e dois nós de saída.



Fonte: Adaptado de Russell e Norvig (2009)

Na Figura 7 é mostrada a curva de aprendizagem para uma rede *perceptron* em dois problemas diferentes e é comparada com o processo de aprendizagem de uma AD. Pode-se notar na figura da esquerda que o *perceptron* aprende a função de forma rápida porque a função utilizada é linearmente separável, enquanto a AD não faz nenhum progresso porque a função é muito difícil, embora não seja impossível, para ser representada como AD. Na figura do lado direito, a solução do problema é facilmente representada como uma AD, mas não é linearmente separável.

Figura 7 - Comparação de desempenho dos perceptrons e das ADs.



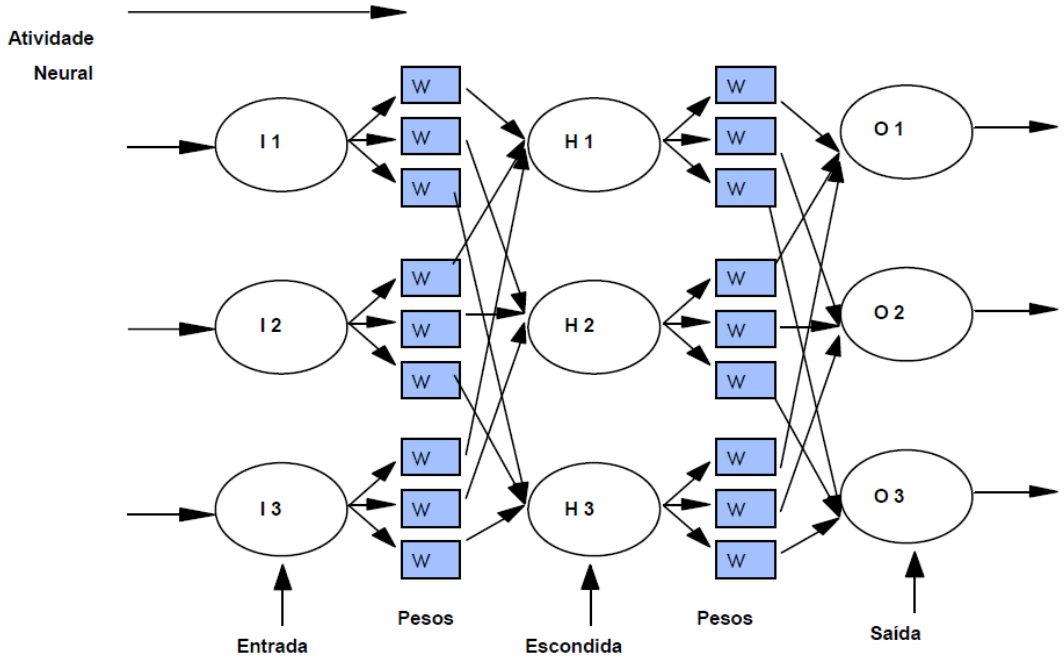
Fonte: Adaptado de Russell e Norvig (2009)

Existem diversos modelos para implementação de uma estrutura de RNA, entre os quais destacam-se: *perceptrons, backpropagation, counterpropagation, Hopfield, Kohonen, bidirecional associative memories (BAM), Self-Organizing Map (SOM), Radius Basis Function (RBF), Least Mean Square (LMS), Multi-Layer Perceptron (MLP).*

A estrutura da RNA utilizada neste trabalho é o MLP. Estas redes de múltiplas camadas distinguem-se das redes de camada simples pelo número de camadas intermediárias, aquelas localizadas entre a camada de entrada e a de saída. Essa arquitetura possui uma ou mais camadas ocultas, que são compostas por neurônios computacionais, também chamados de neurônios ocultos. Segundo Haykin (2008), a função dos neurônios ocultos é intervir entre a camada de entrada externa e a saída da rede de maneira útil. Adicionando-se uma ou mais camadas ocultas, a rede é capaz de extrair estatísticas de ordem elevada. Esta habilidade dos neurônios ocultos é particularmente valiosa quando o tamanho da camada de entrada é grande.

Na Figura 8 ilustra-se uma RNA de múltiplas camadas e cada uma das partes envolvidas no processo de aprendizado.

Figura 8 - Estrutura de uma RNA.



Fonte: Adaptado de Goldschmidt, Passos e Bezerra (2015)

Cada camada do modelo MLP tem uma função específica Russell e Norvig (2009), que é explicada a seguir:

- **Camada de entrada:** É uma camada não-computacional. Nesta camada não há processamento e é responsável pela recepção e propagação das informações de entrada para a camada seguinte.
- **Camadas ocultas ou intermediárias:** Existem uma ou mais camadas ocultas, compostas por nós. São camadas computacionais e efetuam processamento. Nestas camadas são transmitidas as informações por meio das conexões entre as unidades de entrada e saída. Essas conexões guardam os pesos que serão multiplicados pelas entradas, garantindo o conhecimento da rede.
- **Camada de saída:** Camada composta por neurônios computacionais que recebem as informações das camadas ocultas fornecendo a resposta.

A principal complicação que apresentam as redes MLP vem da adição de camadas ocultas da rede. Enquanto que o erro na camada de saída é claro, o erro nas camadas ocultas parece misterioso porque os dados de treinamento não dizem que valores os nós ocultos devem ter. Mas a solução a este problema é retropropagar o erro da camada de saída para as camadas ocultas. O processo de retropropagação emerge diretamente de uma derivação do gradiente de erro geral, como mostrado em Russell e Norvig (2009).

O processo de retropropagação pode ser resumido da seguinte forma:

- Calcular valores Δ para as unidades de saída usando o erro observado.
- A partir da camada de saída, repetir o seguinte para cada camada da rede até que a primeira camada oculta seja alcançada:

  Propagar os valores Δ de volta à camada anterior.

  Atualizar os pesos entre as duas camadas.

Certamente as RNAs são capazes de realizar tarefas muito mais complexas de aprendizagem, embora deva ser dito que é necessário certa quantidade de esforço para obter a estrutura da rede correta e alcançar a convergência para um valor próximo ao ótimo global no espaço de peso.

Até aqui, foi considerado o problema de pesos de aprendizagem, dada uma estrutura de rede fixa. Portanto, faz-se necessário saber como encontrar a melhor estrutura de rede. Se for escolhida uma rede muito grande, ela consegue memorizar todos os exemplos, formando uma tabela grande de pesquisa, mas provavelmente não generalizaria bem as entradas que nunca foram vistas antes. Isto é, como todos os modelos estatísticos, as redes neurais estão sujeitas a super-adaptação quando existem muitos parâmetros no modelo.

Caso fossem consideradas só as redes totalmente conectadas, as únicas escolhas a serem feitas dizem respeito ao número de camadas ocultas e seus tamanhos. A abordagem usual é tentar várias mantendo as melhores. Para isto, é necessário a utilização da técnica de validação cruzada se a pretensão é evitar espreitar o conjunto de teste. Em outras palavras, escolhe-se a arquitetura de rede que oferece a maior previsão de precisão nos conjuntos de validação.

No caso em que são consideradas as redes que não estão totalmente conectadas, é preciso encontrar algum método de busca eficaz através do espaço muito grande de topologias possíveis de conexão. Neste caso, pode ser utilizado o algoritmo de dano cerebral ótimo o qual começa com uma rede totalmente conectada e remove pouco a pouco as conexões dela. Assim que a rede é instruída pela primeira vez, uma abordagem teórica de informação identifica uma seleção ideal de conexões que podem ser descartadas. A rede é, então, reciclada. Se o seu desempenho não diminuir, o processo será repetido. Adicionalmente, além de remover as conexões, também é possível remover as unidades ou nós que não estão contribuindo muito para o resultado.

Tem-se observado que as redes muito grandes fazem bem a generalização, desde que os pesos sejam mantidos pequenos. Essa restrição mantém os valores de ativação na região linear da função sigmoide onde x é próximo de zero. Isto quer dizer, que a rede se comporta como uma função linear, com muito menos parâmetros.

Cabe salientar que as RNAs são frequentemente usadas em casos em que são adequadas saídas múltiplas, além disso, quando o problema de aprendizagem envolve classificação em mais de duas classes.

2.3.2.4 Processo de Treinamento

Treinar uma rede neural significa ajustar sua matriz de pesos de forma que o vetor de saída coincida com um certo valor desejado para cada vetor de entrada. Existem também alguns algoritmos de treinamento que, além de ajustar os pesos, provocam também mudanças na própria estrutura da rede, como a criação ou eliminação de neurônios. O treinamento pode ser de dois tipos: supervisionado ou não supervisionado.

- **Treinamento supervisionado:** exige a disponibilidade de um conjunto de treinamento formado por pares de vetores de entrada e saída, chamados pares de treinamento.
- **Treinamento não supervisionado:** o conjunto de treinamento consiste somente de vetores de entrada.

Uma rede pode ser treinada com três objetivos diferentes:

- **Auto-associação:** após o treinamento com conjunto de vetores, quando submetida a um vetor similiar a um dos exemplos, com algum tipo de perturbação, reconstituir o vetor original.
- **Hetero-associação:** após o treinamento com um conjunto de pares de vetores, quando submetida a um vetor similiar ao primeiro elemento de um par, com algum tipo de perturbação, reconstituir o segundo elemento do par.
- **Detecção de regularidades:** descobrir as regularidades inerentes aos vetores de treinamento e criar padrões para classificá-los de acordo com tais regularidades.

A maioria dos algoritmos de treinamento de RNAs é inspirado, direta ou indiretamente, na lei de Hebb, proposta por Hebb (2008), quem afirma:

> *"A intensidade de uma ligação sináptica entre dois neurônios aumenta se ambos são excitados simultaneamente."*

O algoritmo mais conhecido para treinamento das RNAs é a retropropagação *(back-propagation)*. Este algoritmo pode ser considerado como uma generalização da regra delta para redes de alimentação para frente com mais de duas camadas (RUSSELL; NORVIG, 2009). A retropropagação é um algoritmo de treinamento supervisionado.

Como o algoritmo de retropropagação requer o cálculo do gradiente do vetor de erro, é interessante que a função de ativação seja derivável em todos os pontos. Isto explica o sucesso da função sigmóide como função de ativação, pois ela apresenta esta propriedade. O algoritmo de retropropagação esta conformado pela interação de duas fases, uma fase para frente (forward) e uma fase para trás (backward). Cada uma destas fases é explicada a seguir:

- **Forward**: nesta fase é apresentado à rede cada atributo de entrada. O atributo é recebido pelos neurônios da primeira camada, estando ali é realizada a ponderação do peso associado a suas conexões de entradas correspondentes. Nesta camada cada neurônio aplica sua função de ativação à sua entrada total e gera um valor de saída. Este valor é utilizado como valor de entrada pelos neurônios da seguinte camada. Este processo é repetido até que os neurônios da camada de saída gerem seu valor de saída, este valor é comparado ao valor *alvo* para a saída desse neurônio. Neste ponto é determinado o erro cometido pela rede para o *alvo* esperado, medindo a diferença entre os valores de saída gerados e os desejados para cada neurônio da camada de saída.

- **Backward:** nesta fase é utilizado o valor do erro de cada neurônio da camada de saída para ajustar seus pesos de entrada. O ajuste é realizado desde a camada de saída até a primeira camada intermediária. Dado que os valores dos erros só são conhecidos para os neurônios da camada de saída, é necessário estimar o erro para os neurônios das camadas intermediárias.

O algoritmo de retropropagação utiliza os erros observados nos neurônios da camada posterior para assim conseguir estimar o erro dos neurônios das camadas intermediárias. O erro de um neurônio de uma dada camada intermediária é estimado como a soma dos erros dos neurônios da camada seguinte, cujos terminais de entradas estão conectados a ele, ponderados pelo valor do peso associado a essas conexões (GAMA et al., 2011).

Além do algoritmo de retropropagação, existem outros algoritmos de treinamento de RNAs, os mais conhecidos na literatura especializada são:

- Contrapropagação (counterpropagation).
- Aprendizado competitivo, utilizado nas redes de Kohonen.
- Algoritmos Genéticos, entre outros que não foram considerados nesta tese.

Para o desenvolvimento desta tese foi utilizado o algoritmo de treinamento de retropropagação, a escolha foi feita considerando a eficiência computacional deste algoritmo apresentada em pesquisas encontradas na literatura.

2.3.3 Paradigma estatístico

Neste paradigma, métodos estatísticos, em geral, paramétricos, são utilizados para encontrar boas aproximações do modelo de conhecimento que esteja sendo induzido.

2.3.3.1 Máquinas de Vetor de Suporte

A teoria de MVS foi desenvolvida por Vapnik (1998) baseado na ideia de minimização do risco estrutural, o que tem apresentado um ótimo desempenho. Segundo Russell e Norvig (2009), a estrutura MVS é atualmente a abordagem pré-fabricada mais popular para aprendizagem supervisionada, devido as seguintes propriedades:

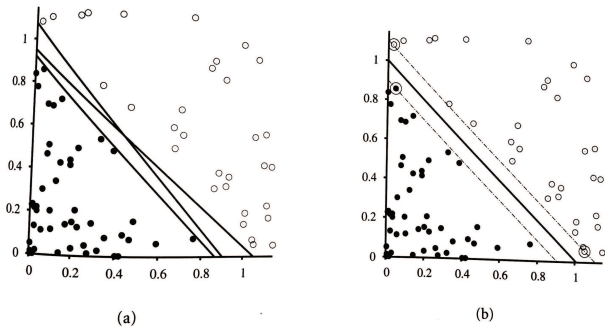
- As MVS constroem um separador de margem máxima, isto é, um limite de decisão com a maior distância possível a pontos de exemplo. O que permite uma correta generalização.
- As MVS criam uma separação linear em hiperplano, mas, têm a capacidade de incorporar os dados em um espaço de dimensão superior, usando assim o chamado *kernel*. Muitas vezes, os dados que não são separáveis linearmente no espaço de entrada original são facilmente separáveis em um espaço de dimensão superior. O separador linear de dimensão superior é realmente não linear no espaço original. Isto quer dizer, que o espaço de hipótese é expandido em relação aos métodos que usam representações estritamente lineares.
- As MVS são um método não paramétrico, isto é, elas mantêm exemplos de treinamento e podem precisar armazenar todos eles. Por outro lado, na prática, acabam mantendo apenas uma pequena fração do número de exemplos, às vezes apenas uma constante do número de dimensões. Assim, as MVSs combinam as vantagens de modelos não paramétricos e paramétricos: eles têm a flexibilidade para representar funções complexas, mas são resistentes à superadaptação.

A ideia das MVS, é que alguns exemplos são mais importantes que os outros e dar mais atenção a eles pode levar a melhor generalização.

As MVSs fazem parte da família de classificadores lineares, pois realizam o mapeamento dos pontos de entrada num espaço de características de uma dimensão maior, para depois achar o hiperplano que os separe e maximize a margem entre as classes. Para separar duas classes é utilizado o chamado separador de margem máxima. A margem é a largura da zona delimitada, e o separador é definido como um conjunto de pontos $x : w * x + b = 0$. Assim, um hiperplano separador (w, b) não só deve classificar os dados corretamente, mas também deve fazer as margens (γ) tão grande quanto possível, Russell e Norvig (2009).

Na Figura 9 é mostrada uma MVS, no lado esquerdo, que corresponde à Figura 9(a) podem ser observados duas classes de pontos, alguns são pretos e outros são círculos brancos, e três separadores lineares candidatos. No lado direito, que corresponde a Figura 9(b) se tem o separador de margem máxima, que está no ponto médio da margem, isto é, a área entre as linhas tracejadas. O suporte vetorial representado na Figura 9 como pontos com grandes círculos, é o exemplo mais próximo do separador.

Figura 9 - MVS (a) Duas classes de pontos. (b) Separador de margem máxima



Fonte: Adaptado de Russell e Norvig (2009)

A formulação matemática das MVSs varia dependendo da natureza dos dados, isto é, existe uma formulação para os casos lineares e uma formulação para casos não lineares.

- **MVS com margens rígidas:** este tipo de MVS define fronteiras lineares a partir de dados linearmente separáveis. Seja X um conjunto de treinamento com n objetos $x_i \in X$ e seus respetivos rótulos $y_i \in Y$, em que X constitui o espaço de entradas e $Y = \{-1, +1\}$ são as classes. Pode-se afirmar que X é linearmente separável se é possível separar os objetos das classes $+1$ e -1 por um hiperplano. Este tipo de classificadores que separam os dados por meio de um hiperplano são denominados lineares, Gama et al. (2011).

- **MVS com margens suaves:** são utilizadas na resolução de problemas que apresentam ruídos nos dados, e *outliers* nos objetos, e em problemas não lineares. Este tipo de MVS suaviza as margens do classificador linear, permitindo que alguns objetos permaneçam entre os hiperplanos H_1 e H_2 e também a ocorrência de alguns erros de classificação, Gama et al. (2011).

- **MVSs não lineares:** são utilizadas em problemas em que não é possível dividir os dados de treinamento por um hiperplano, levando ao uso de uma fronteira curva para realizar a separação das classes, Gama et al. (2011).

A representação alternativa para os casos não lineares é chamada de representação dual, em

que a solução ótima é encontrada resolvendo:

$$\max_{\alpha} \sum_{j} \alpha_j - \frac{1}{2} \sum_{j,k} y_j y_k \alpha_j \alpha_k < x_j * x_k > \quad (1)$$

sujeito a: $\alpha_j \geq 0$, $\sum_j \alpha_j y_j = 0$

Esse é um problema de otimização de programação quadrática. Uma vez encontrado o vetor, pode-se voltar a w com a equação $w = \sum_j \alpha_j x_j$ ou pode-se ficar com a representação dual. A equação (1), tem três propriedades importantes:

- A expressão é convexa, isto é, tem um único máximo global que pode ser encontrado de forma eficiente.
- Os dados inserem a expressão apenas na forma de produtos escalares de pares de pontos.
- Os pesos j associados a cada ponto de dados são zero, exceto pelos vetores de suporte, que são os pontos mais próximos do separador.

Como, normalmente, há muito menos vetores de suporte que exemplos, as MVSs obtêm algumas das vantagens dos modelos paramétricos.

No entanto, os problemas do mundo real geralmente não são lineares separáveis e é nesse ponto onde o método do kernel é introduzido (MULLER et al., 2001). A função kernel pode ser aplicada a pares de dados de entrada para avaliar produtos escalares em algum espaço característico. Assim, pode-se aprender no espaço de dimensão superior, mas, calcula-se somente as funções kernel, em vez de, uma lista completa de características para cada ponto de dados.

O teorema de Mercel (1909) diz que qualquer função kernel corresponde a um espaço característico, que pode ser muito grande, mesmo para kernels que parecem inócuos. Por exemplo, o kernel polinomial $k(x_j, x_k) = (1 + x_j * x_k)^d$ corresponde a um espaço caraterístico cuja dimensão é exponencial em d.

Segundo Russell e Norvig (2009), o truque do kernel inteligente é ligado ao kernel escolhido na equação (1), no qual, podem ser encontrados ótimos separadores lineares de forma eficiente em espaços característicos com bilhões de dimensões. Os separadores lineares resultantes, quando mapeados de volta ao espaço de entrada original, podem corresponder a fronteiras de decisão arbitrariamente não lineares, sinuosas entre os exemplos.

O kernel polinomial, representa a expansão a todas as combinações de monômios de ordem d. O kernel polinomial evita o problema de nulidade da hessiana (matriz utilizada em técnicas de otimização envolvendo a função kernel). Esta função é simétrica e, re-escrita em coordenadas espaciais, descreve o produto interno no espaço de características que contém todos os produtos $x_{i1}, \ldots, x_{ip}$ até o grau p.

Apesar do espaço de características de alta dimensionalidade (polinómios de grau d no espaço de entrada tem n^p parâmetros livres), a estimativa da dimensão do subconjunto de polinómios que solucionam problemas práticos (com base em um dado conjunto de treinamento) pode ser baixa segundo o Teorema 10.3 apresentado por Vapnik (1998). Se a esperança desta estimativa é baixa, então a esperança da probabilidade de erro é pequena segundo o Teorema 10.58, apresentado em Vapnik (1998).

O caso de dados inerentemente ruidosos pode não precisar um separador linear em algum espaço de dimensão superior. Em vez disso, é preferível uma superfície de decisão em um espaço de dimensão inferior que não separe as classes claramente, mas reflita a realidade dos estados ruidosos. Isto é possível mediante o uso do classificador de margem suave, que permite que os exemplos caiam no lado errado da fronteira de decisão, mas lhes atribui a penalidade proporcional à distância necessária para movê-los de volta ao lado correto, Russell e Norvig (2009).

O método de kernel pode ser aplicado não só a algoritmos de aprendizagem que encontram os melhores separadores lineares, mas também a qualquer outro algoritmo que possa ser reformulado para funcionar somente com produtos escalares de pares de pontos de dados.

As MVSs têm sido desenvolvidas como uma técnica robusta para classificação e regresão aplicado a grandes conjuntos de dados complexos e com ruidos; isto é, com variaveis inherentes ao modelo que para outras técnicas aumentam a possibilidade de erro nos resultados, isto porque resultam difíceis de quantificar e observar. Maiores detalhes sobre MVS e método kernel podem ser encontradas em Vapnik (1998), Shawe-Taylor e Cristianini (2000) e Tan e Wang (2004).

2.4 MINERAÇÃO DE DADOS

Considerando as dificuldades que aparecem ao realizar a análise de grandes quantidades de dados sem o uso de ferramentas computacionais apropriadas, foram criadas metodologias que ajudam nas tarefas de analisar, interpretar, e relacionar esses dados, para que se possa elaborar e selecionar estratégias de ação em cada domínio.

Nesse contexto, cabe esclarecer que a expressão *Mineração de Dados*, é realmente, uma das etapas do processo de Descoberta de Conhecimento em Bases de Dados (DCBD). Nos últimos anos, esta área vem despertando grande interesse junto as comunidades científicas e industriais (HAN; MICHELINE; JIAN, 2012).

Para termos uma melhor compreensão do que é a MD é necessário estudar suas bases, saber o que existe por trás dela, para isto, é preciso conhecer alguns termos, mostrados adiante:

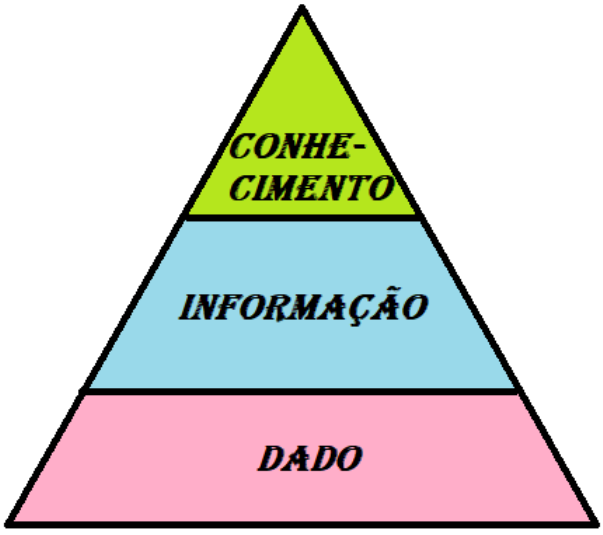
2.4.1 Sistema Baseado em Conhecimento

Um Sistema Baseado em Conhecimento (SBC) pode ser definido como sendo um sistema computacional que representa e utiliza conhecimento para resolução de problemas. Segundo (REZENDE; PUGLIESE; AO, 2003), simples bases de dados não atendem mais à demanda das empresas, portanto, a manipulação da informação é um fator de grande importância para as empresas que estejam interessadas em se diferenciar no mercado competitivo.

Segundo o apresentado por Han, Micheline e Jian (2012), fazer referência às tecnologias de extração de conhecimento, pode ser subentendido como o uso de ferramentas computacionais capazes de realizar consultas e análises complexas em grandes volumes de dados, a partir da utilização de processos de extração de informações implícitas.

Estes dados, relações, informações genéricas, relevantes e previamente desconhecidos podem ser extraídos a partir da formulação prévia de hipóteses. Na Figura 10, ilustra-se a hierarquia entre dados, informação e conhecimento. Estes três conceitos são totalmente diferentes, e para se ter um melhor entendimento do que realmente é a MD é necessário destacar as diferenças entre eles.

Figura 10 - Hierarquia entre Dado, Informação e Conhecimento.



Fonte: Adaptado de Goldschmidt, Passos e Bezerra (2015)

Os dados são cadeias de símbolos e não possuem semântica, podem ser interpretados como itens elementares, captados e armazenados por recursos tecnológicos. O propósito dos dados é expressar fatos reais, de forma tal, que possam ser tratados computacionalmente. Por exemplo, uma cadeia de símbolos como: "313.972"corresponde a um dado. Quando este valor é identificado como sendo a quantidade de usuários de uma empresa distribuidora de energia elétrica, se passa a ter uma informação.

A informação representa os dados com significados e contextos bem definidos.

O conhecimento, é um conceito que corresponde a um padrão ou conjunto de padrões cuja formulação pode envolver e relacionar dados e informações. O conhecimento pode-se encontrar representado na forma de uma regra condicional.

Uma vez conhecida a hierarquia entre os dados, a informação e o conhecimento, pode-se dizer que existem dificuldades e limitações no uso da metodologia de SBC, as quais dependem da base de conhecimento, ou seja, um SBC fica restrito ao conhecimento existente na sua base de dados e ao conhecimento que possa ser adicionado. Segundo isto, um dos maiores desafios na construção de um SBC é a aquisição do conhecimento, a sua manutenção bem como a dificuldade de se avaliar ou prever como será o desempenho do sistema para tratar os casos reais.

Goldschmidt, Passos e Bezerra (2015) afirma que a informação e conhecimento constituem, em geral, a base para se tomar decisões em diversos cenários. Considerando que o conhecimento não pode ser obtido de bases de dados simplesmente utilizando recursos tradicionais, se torna necessário o uso de tecnologias especificas para a realização da abstração de conhecimento. A busca de novos conhecimentos a partir dos dados é o tema central dos sistemas de DCBD que são mostrados a seguir.

2.4.2 Sistemas de Descoberta do Conhecimento em Bases de Dados

Os sistemas DCBD, também chamados como *Knowledge Discovery in Databases* (KDD), podem ser vistos como processos de descoberta de novas correlações, padrões e tendências significativas por meio da análise minuciosa de grandes conjuntos de dados. Estes processos baseiam-se em tecnologias de reconhecimento utilizando padrões e técnicas estatísticas e matemáticas (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

As técnicas utilizadas no processo de DCBD tiveram suas origens em diferentes áreas de pesquisa. Antes da popularização dos sistemas de DCBD e da MD, pesquisadores em áreas como IC, AM, reconhecimento de padrões, estatística, entre outras, já trabalhavam sobre os mesmos tipos de problemas sem existir nenhuma integração entre uma e outra técnica. Os sistemas de DCBD reúnem todas estas disciplinas sob a premissa de que bancos de dados guardam mais informação do que os dados neles armazenados, (HAN; MICHELINE; JIAN, 2012).

A complexidade do processo de DCBD está na dificuldade em perceber e interpretar adequadamente inúmeros fatos observáveis durante a realização do processo, e na dificuldade em conjugar dinamicamente tais interpretações de forma a decidir que ações devem ser realizadas em cada caso (GOLDSCHMIDT; PASSOS, 2005).

De acordo com a definição de DCBD, pode-se afirmar que um dos propósito de realizar

o processo de DCBD é identificar padrões. Um padrão é dito compreensível quando a sua representação do conhecimento pode ser interpretada por seres humanos.

Os padrões extraídos no processo de DCBD podem ser classificados como preditivos e descritivos, descritos adiante.

- **Padrões preditivos:** são construídos com o intuito de resolver problemas específicos para prever os valores de um ou mais atributos, em função dos valores de outros atributos. São avaliados pelo julgamento de quão efetivos eles são na predição. Quanto mais próximas as predições forem dos valores reais, mais efetivos serão os padrões avaliados.
- **Padrões descritivos:** busca apresentar informação interessante que um especialista do domínio da aplicação possa ainda não conhecer. São avaliados não só em função da sua natureza subjetiva, mas também em virtude da sua contribuição.

Os resultados do processo de DCBD devem conter padrões extraídos da base de dados que sejam válidos, novos, úteis, interessantes e inteligíveis. Sendo assim, é valido afirmar que a DCBD está determinada por um conjunto de atividades contínuas que compartilham o conhecimento descoberto, a partir de bases de dados muito grandes.

Para realizar o processo de DCBD é necessária a utilização de ferramentas e técnicas de análise de dados. Porém, antes de realizar qualquer etapa do processo de DCBD, é necessário entender os objetivos do problema e ter domínio do mesmo, para que seja possível a seleção dos dados relevantes e que posteriormente possa ser selecionado e recolhido o conjunto de dados, ou variáveis necessárias para resolver algum tipo de problema. Deve-se ter em consideração que a qualidade da preparação dos dados pode aproximar ou distanciar os resultados do processo de MD da solução ideal (ENGELS; THEUSINGER, 1998).

2.4.3 Etapas do processo de DCBD

O conjunto de atividades contínuas que compartilham o conhecimento descoberto está composto de etapas operacionais ilustradas na Figura 11.

Figura 11 - Etapas do Processo de DCBD.

Fonte: Adaptado de Goldschmidt, Passos e Bezerra (2015)

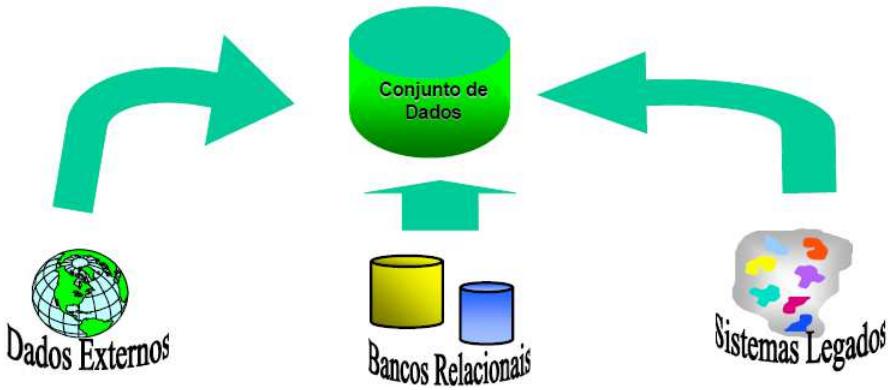
A etapa de pré-processamento é a encarregada das funções relacionadas com a captação, a organização e o tratamento de dados e tem como objetivo a preparação dos dados para o algoritmo da etapa seguinte, a MD. Na etapa de MD é realizada a busca por informações e conhecimentos úteis para à aplicação de DCBD. Na etapa de pós-processamento é realizado o tratamento das informações e do conhecimento obtido na MD (GOLDSCHMIDT; PASSOS; BEZERRA, 2015). Para uma melhor compreensão, cada etapa do processo de DCBD é descrita adiante.

2.4.3.1 Etapa de pré-processamento

A etapa de **pré-processamento** está composta por todas as atividades relacionadas com a captação, a organização e o tratamento dos dados. O objetivo desta etapa é a preparação dos dados para a aplicação dos algoritmos na etapa de MD. As funções principais da etapa de pré-processamento dos dados são: seleção, limpeza, codificação e enriquecimento dos dados. Uma breve descrição é realizada abaixo:

- **Seleção de dados ou redução de dados:** esta função destaca-se pela identificação do subconjunto de bases de dados existentes que devem ser considerados durante o processo de DCBD. A seleção de dados pode ser realizada considerando dois enfoques diferentes: a seleção de atributos ou a seleção de registros. Na Figura 12 é ilustrada a extração dos dados de diversas fontes. Posteriormente, estes dados passam para um banco de dados onde são analisados e é realizada a escolha dentre os dados de um novo conjunto de dados.

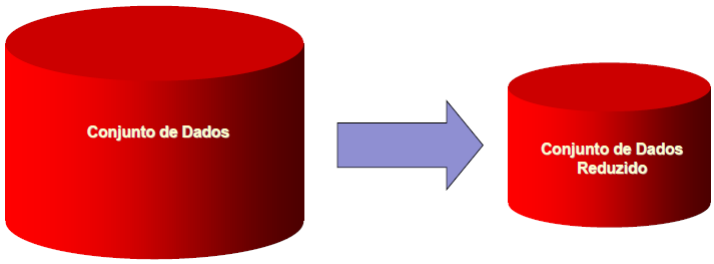
Figura 12 - Seleção de dados.



Fonte: Adaptado de Goldschmidt, Passos e Bezerra (2015)

No contexto da seleção de registros, é possível que nem todo o conjunto de dados possa ser usado por algum motivo (*e.g.*, ausência de valores de atributos em alguns registros, valores inconsistentes, etc.). Nesse caso, para compor o conjunto de dados a ser analisados é selecionado um subconjunto de registros. Na realização deste processo é reduzido o tamanho da base de dados inicialmente obtida, como ilustrado na Figura 13.

Figura 13 - Redução de dados.



Fonte: Adaptado de Goldschmidt, Passos e Bezerra (2015)

A seleção de dados ou redução de dados é realizada com o intuito de eliminar atributos redundantes ou irrelevantes, obter uma representação reduzida em volume mas que produz resultados de análise idênticos ou similares aos obtidos com o conjunto completo de atributos e melhorar o desempenho dos modelos de aprendizado (HAN; MICHELINE; JIAN, 2012).

- **Limpeza de Dados:** esta função envolve toda e qualquer atividade realizada sobre os dados selecionados, visando garantir a complexidade, veracidade e integridade dos fatos representados. Em caso de existir informações inconsistentes, erróneas ou ausentes, as mesmas devem ser corrigidas, garantindo assim, a qualidade dos modelos de conhecimento a serem extraídos ao final do processo de DCBD (GOLDSCHMIDT; PASSOS;

BEZERRA, 2015).

Cabe ressaltar que, em aplicações práticas, é comum a existência de dados que estejam incompletos, ruidosos ou inconsistentes. Entenda-se como dados incompletos aquelas informações ausentes para determinados atributos ou dados pouco detalhados. Dados ruidosos são aqueles dados errados ou que contém valores considerados divergentes (outliers). Por último encontram-se os dados inconsistentes que são aqueles dados que contêm algum tipo de discrepância semântica entre si.

O processo de limpeza de dados envolve varias atividades, entre elas têm-se: verificação da consistência das informações, correção de possíveis erros, complementação ou eliminação de valores desconhecidos, eliminação de valores não pertencentes ao domínio, entre outras (GOLDSCHMIDT; PASSOS; BEZERRA, 2015).

- **Codificação dos dados:** nesta função, deve ser garantido que os dados podem ser usados como entrada dos algoritmos de MD. Para isto é necessário codificar os dados numericamente antes de ser submetidos a determinados algoritmos de MD. A codificação pode ser Numérica-Categórica ou Categórica-Numérica. Na codificação do tipo Numérica-Categórica também conhecida como mapeamento Direto é realizada uma simples substituição dos valores numéricos por valores categóricos ou intervalos.

  Por outro lado, a codificação Categórica-Numérica também conhecida como Mapeamento em Intervalos ou Discretização é uma técnica que envolve métodos que dividem o domínio de uma variável numérica em intervalos. Estes intervalos podem ter comprimentos definidos pelo usuário, neste caso, o analista de dados define o número de intervalos e escolhe o tamanho de cada um deles; ou pode ser realizada uma divisão em intervalos de igual comprimento, neste caso, o analista de dados define somente o número de intervalos, o comprimento destes intervalos é calculado a partir do maior e menor valor do domínio do atributo (GOLDSCHMIDT; PASSOS; BEZERRA, 2015).

- **Normalização:** Segundo (GOLDSCHMIDT; PASSOS; BEZERRA, 2015) *"consiste em ajustar a escala dos valores de cada atributo de forma que os valores fiquem em pequenos intervalos, tais como de -1 a 1, ou de 0 a 1. Este ajuste evita que alguns atributos, por apresentarem uma escala de valores maiores que outros, influenciem de forma tendenciosa em determinados métodos de Mineração de Dados."*

- **Enriquecimento dos dados:** Segundo (GOLDSCHMIDT; PASSOS; BEZERRA, 2015) *"este processo consiste em conseguir de alguma forma mais informação, que possa ser unida aos registros existentes, agregando novos fatos aos dados existentes e oferecendo mais opções para o processo de DCBD."*

- **Construção de Atributos:** Segundo (GOLDSCHMIDT; PASSOS; BEZERRA, 2015) *"esta operação consiste em gerar novos atributos a partir de atributos existentes. Os novos atributos são denominados atributos derivados."*

- **Correção de Prevalência:** Segundo (GOLDSCHMIDT; PASSOS; BEZERRA, 2015) *"esta operação é muitas vezes necessária em tarefas de classificação. Consiste em corrigir um eventual desequilíbrio na distribuição de registros com determinadas características."*

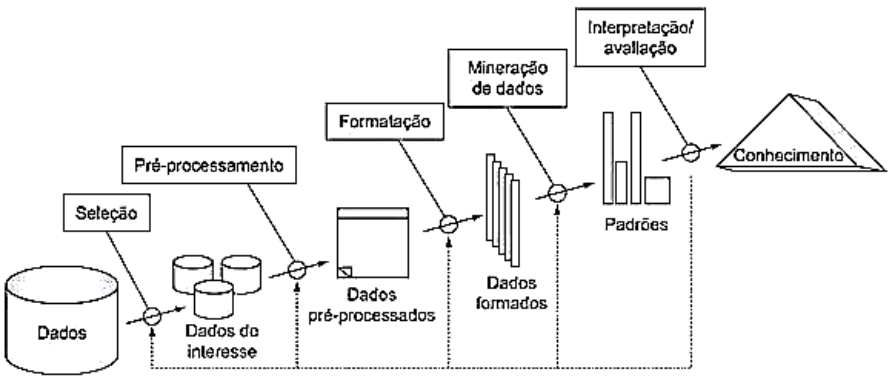
- **Partição dos Dados:** Segundo (GOLDSCHMIDT; PASSOS; BEZERRA, 2015) *"esta função consiste em separar os dados em dados para treinamento (abstração do modelo de conhecimento) e em dados para testes (avaliação do modelo gerado). "*

2.4.3.2 Etapa de Mineração de Dados

Embora existam muitas definições sobre o que é MD do ponto de vista de diferentes autores, talvez, a definição que se destaca por ser mais completa é a de Fayyad, Piatetsky-Shapiro e Smyth (1996): *"...o processo não trivial de identificar, em dados, padrões válidos, novos, potencialmente úteis e ultimamente compreensíveis"*.

A MD é uma das etapas realizadas durante o processo de DCBD, (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996), assim como ilustrado na Figura 14. O uso da MD torna possível a transformação de dados em informação e posteriormente em conhecimento que pode ser relevante em processos de tomada de decisões. Segundo Silva (2004) é necessário determinar qual é a relevância do conhecimento em um processo de tomada de decisões, bem como eventuais impactos que este conhecimento tem nas medidas, ou soluções concretizadas para a empresa ou pesquisa.

Figura 14 - Etapas do processo de DCBD.

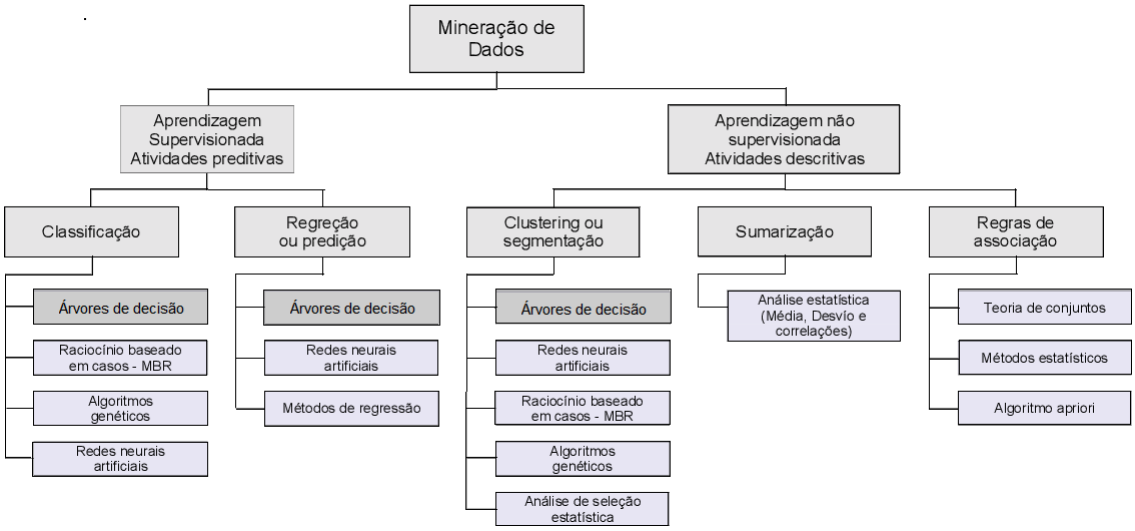


Fonte: Adaptado de Fayyad, Piatetsky-Shapiro e Smyth (1996)

A etapa de MD abrange o processo de busca por conhecimentos úteis no contexto da aplicação de DCBD. Na literatura alguns autores se referem à MD e a DCBD como sendo sinônimos, mas realmente a MD é a etapa principal do processo de DCBD.

Nesta etapa são definidos os algoritmos e as técnicas a serem utilizadas na resolução do problema que está sendo tratado. Alguns exemplos de técnicas que podem ser usadas nesta etapa são: RNA (HAYKIN, 2008), MVS (PLATT, 1998), ADs (QUINLAN, 1996), entre outros modelos estatísticos e probabilísticos. A escolha do algoritmo e da técnica que vão ser usadas nesta etapa depende do tipo de tarefa de DCBD que vai ser realizada assim como ilustrado na Figura 15.

Figura 15 - Descrição do processo de MD.



Fonte: Adaptado de Delgado (2010)

Cada um destes algoritmos são fundamentados em técnicas que procuram, segundo determinados paradigmas de aprendizado, explorar os dados visando produzir modelos de conhecimento. O algoritmo de MD esta diretamente relacionado com a forma de representação do conhecimento em um modelo de conhecimento.

Deve ser considerado que, todo conjunto de dados no processo de DCBD corresponde a uma base de fatos armazenados que devem ser interpretados como um conjunto de pontos em um hiperespaço de dimensão *K*. Esta dimensão está determinada pelo número de atributos do conjunto de dados em análise.

Além disso, todo processo de DCBD deve ter um objetivo claramente definido, pois, este objetivo compreende as definições das tarefas do processo de DCBD a ser executadas e da expectativa dos conhecedores do domínio da aplicação, com relação ao modelo de conhecimento a ser gerado. Quer dizer que o especialista em DCBD tem condições de delinear que tipos de

padrões devem ser abstraídos a partir das bases de dados (HAN; MICHELINE; JIAN, 2012).

Uma vantagem que tem a MD é que para suportar todo o processo de extração de conhecimento dos dados, esta se apoia na modelagem matemática para auxiliar a identificação de padrões nos dados observados. Porém, se os resultados da MD representam ou não conhecimento em um determinado domínio, depende da técnica e do algoritmo utilizado.

Os algoritmos de MD selecionados devem percorrer um espaço de padrões, (também conhecido como espaço de hipóteses ou espaço de busca), em busca de padrões que atendam as condições relacionadas ao modelo de conhecimento desejado. Para concluir, pode-se dizer que padrões podem ser ordenados com relação a alguma medida de interesse que expresse sua adequação ao problema que está sendo tratado. Neste caso, pode ser utilizada a acurácia mínima de um modelo de conhecimento para medir a expectativa sobre um determinado modelo, considerando que a acurácia é uma medida que expressa a capacidade de o modelo de conhecimento produzir uma resposta correta diante de casos a ele apresentados, (GOLDSCHMIDT; PASSOS; BEZERRA, 2015).

Outro conceito relevante sobre MD se refere à capacidade de aprendizado a partir de exemplos que determinados algoritmos têm. Segundo Han, Micheline e Jian (2012), os algoritmos aprendem os relacionamentos eventualmente existentes entre os dados, retratando o resultado deste aprendizado nos modelos de conhecimento gerados. As principais abordagens de aprendizado aplicadas em MD são: o aprendizado supervisionado e o aprendizado não-supervisionado, que já foram definidos na seção 2.3.

Em Silva e Robin (2003) é apresentado um estudo no qual se descreve o uso da MD em um ambiente *WEB*. Neste estudo foram aplicadas técnicas de DCBD. Para analisar os dados, foi utilizada uma ferramenta muito empregada nas pesquisas atualmente desenvolvidas na área de MD. Esta ferramenta é chamada de WEKA e permite realizar o processo de MD aplicando diversos algoritmos de AM. Varias aplicações da ferramenta WEKA para resolver problemas tanto reais quanto acadêmicos utilizando diferentes algoritmos de AM são apresentadas em (WITTEN et al., 2016).

2.4.3.3 Etapa de pós-processamento

A etapa de pós-processamento envolve a visualização, a análise e a interpretação do modelo de conhecimento gerado na etapa de MD. É realizado o tratamento do conhecimento obtido com o intuito de facilitar a interpretação e fazer a avaliação dos resultados obtidos. As principais funções que devem ser realizadas nesta etapa são: elaboração e organização do conhecimento obtido, onde podem se incluir gráficos, diagramas, ou relatórios demonstrativos. Para alguns pesquisadores esta etapa é desnecessária (GOLDSCHMIDT; PASSOS; BEZERRA, 2015).

2.5 CONSIDERAÇÕES FINAIS

Este capítulo apresentou as diferentes técnicas de IC utilizadas no desenvolvimento desta tese. Entre elas são destacados os algoritmos de AM: ADs, RNAs e MVS. Estes algoritmos foram escolhidos para realizar a pesquisa apresentada nesta tese porque fazem uso de um processo de otimização para obter modelos preditivos a partir de um conjunto de dados, além disso, apresentam as seguintes características:

- As ADs são um dos algoritmos mais comummente usados devido a suas características. Alguns dos pontos mais relevantes referenciados na literatura são: a flexibilidade, a robustez, a seleção de atributos, a interpretabilidade e a eficiência. As ADs são uma aproximação para à tomada de decisões multietapas, as quais são amplamente usadas em muitas aplicações. Uma das principais características das ADs é a capacidade de dividir uma decisão complexa em um conjunto de decisões mais simples, fornecendo soluções acertadas similares aos objetivos desejados (MEI; ROVNYAK, 2004).

- Na literatura pode ser constatado que as RNAs mais empregadas na prática são as redes MLP, este tipo de rede é utilizada em conjunto com o algoritmo de treinamento mais popular que é o *back-propagation*. As principais características deste algoritmo são: a) a busca minimizar o erro quadrático entre os atributos desejados e os produzidos pela rede, e b) é um algoritmo baseado no método de otimização gradiente descendente. As RNAs também possuem características que as tornaram populares dentro das técnicas de AM, por exemplo: capacidade de adquirir, armazenar e utilizar conhecimento experimental, capacidade de generalização, tolerância a falhas e ruídos. Estas características fazem com que as RNAs apresentem um bom desempenho.

- Os princípios de uma MVS são baseados na teoria de aprendizado estatístico introduzido por Vapnik (1998). As MVS caracterizam-se pela sua capacidade de generalização, robustez diante de problemas de grande porte, e pela convexidade do problema de otimização formulado em seu treinamento. Em função dos dados de treinamento as MVS recorrem a um problema de otimização quadrático que busca maximizar a margem de separação entre as classes. O objetivo principal das MVSs é construir uma função de decisão ótima, que, a partir de um conjunto de dados com entradas e saídas esperadas, classifique novas entradas e minimize o erro de classificação.

- Uma MVS é capaz de apresentar uma única solução global, enquanto as RNAs MLP podem possuir mínimos locais na sua função de erro minimizada, o que levaria a uma possível convergência do método gradiente para um deles.

- As técnicas de RNAs e MVS são consideradas "caixas pretas", dado que o conhecimento extraído dos dados se encontra codificado em equações de difícil interpretação, em contra-

partida são utilizados algoritmos de ADs que permitem interpretar com maior facilidade os resultados obtidos.

- De forma geral, as ADs, as RNAs e as MVS apresentam bom desempenho preditivo em tarefas de classificação e regressão, destacando-se entre as técnicas mais utilizadas na solução de problemas que requerem alta precisão. Estas técnicas apresentam como característica em comum que são tolerantes a ruídos. O uso das RNAs, MVS e ADs para resolver problemas encontrados nos sistemas de energia elétrica é uma área de crescente interesse Haykin (2008).

3 APLICAÇÃO DE SIS NO CONTROLE CENTRALIZADO VOLT-VAR EM MODERNOS SDEE

Neste capítulo é apresentada a aplicação dos SIs como solução ao problema de controle centralizado Volt-VAr em modernos SDEE. Este capítulo está organizado da seguinte forma: na Seção 3.1 é realizada uma introdução ao problema, e são apresentados os diferentes métodos de solução encontrados na literatura para sua solução. Na Seção 3.2, é apresentada a metodologia de solução proposta. Na Seção 3.3 apresenta-se o caso de estudo com seus resultados. Por último, Na Seção 3.4 apresentam-se as conclusões do capítulo.

3.1 INTRODUÇÃO

Devido à preocupação dos SDEE em serem mais flexíveis aos avanços tecnológicos, ao aumento nos preços dos combustíveis, e ao interesse na conservação do meio ambiente, incentiva-se o desenvolvimento da geração distribuída, sistemas de armazenamento de energia, programas de resposta à demanda, e tecnologias de medição sincronizada para resolver os problemas relacionados com este tipo de variáveis, como componentes essenciais para atingir as *redes inteligentes*.

Dentro do contexto das *redes inteligentes*, um dos problemas existentes nos SDEE é o controle de magnitude de tensão e de potência reativa (Volt-VAr), em que se busca determinar o ajuste ótimo de um conjunto de variáveis de controle existentes na rede com o objetivo de garantir uma adequada operação do SDEE. Entre as principais variáveis de controle podemos destacar a geração de potência ativa e reativa dos geradores distribuidos (GDs), o número de módulos em operação dos bancos de capacitores (BCs), e o número de passos do tap para os transformadores com (OLTC) e para os reguladores de tensão (RTs). Dada sua importância, e de acordo com a filosofia das *smart grids*, existe grande interesse em desenvolver mecanismos que permitam realizar este controle de forma automatizada (KERSTING, 2007).

Neste trabalho são utilizadas três metodologias de SIs para determinar o controle centralizado de magnitude de tensão e potência reativa (Volt-VAr) em tempo real dos SDEE usando medições elétricas. As metodologias de SIs consideradas são ADs, RNA e MVS. Através delas definem-se os ajustes mais apropriados da geração de potência ativa e reativa dos GDs, o número de módulos em operação dos BC, e o número de passos do tap para os transformadores com comutação de tap sob carga e RTs. As medidas elétricas consideradas são as magnitudes de tensão nos nós, as magnitudes de fluxo de corrente nos circuitos e as injeções de potência ativa e reativa nos nós em diferentes pontos da rede. Os SIs são treinados a partir das medi-

ções disponíveis e ações registradas no centro de controle do sistema. Entretanto, um modelo matemático de programação linear inteiro misto (MPLIM) foi usado para simular este conjunto de medições e ações de controle. Os testes foram realizados em um sistema de 42 nós com o objetivo de mostrar a eficiência e a robustez da técnica de solução proposta quando comparada com os resultados do modelo matemático, obtendo resultados satisfatórios por parte dos SIs.

O controle Volt-VAr tem sido amplamente estudado considerando somente como variáveis de controle os taps dos OLTC, RTs e os módulos dos BCs (ARAÚJO; MEIRA; ALMEIDA, 2013; BRADY; DAI; BAGHZOUZ, 2003; GONÇALVES; ALVES; RIDER, 2012; GONÇALVES et al., 2013; LIANG; CHENG, 2001; LIANG; WANG, 2003; PARK; NAM; PARK, 2007; VIAWAN; KARLSSON, 2008). Em Park, Nam e Park (2007), os taps dos RTs são controlados de modo cooperativo em tempo real, após a programação do despacho ótimo dos modulos dos BCs, usando algoritmos genéticos. Em Liang e Cheng (2001), é usado um algoritmo de programação dinâmica para determinar o despacho ótimo dos taps dos RTs e dos módulos dos BCs com o objetivo de minimizar as perdas de potência ativa ou o desvio da magnitude de tensão do sistema. Tanto em Park, Nam e Park (2007), quanto em Liang e Cheng (2001), o impacto da injeção de potência ativa e reativa dos GDs no controle Volt-VAr é desconsiderado. Em Brady, Dai e Baghzouz (2003), é explicado que o controle convencional dos pontos de operação dos módulos dos BCs em redes radiais precisa ser revisado para incluir os GDs nos alimentadores. Em Viawan e Karlsson (2008), a coordenação dos taps dos RTs e dos módulos dos BCs é proposta com o objetivo de minimizar as perdas ativas no sistema considerando o impacto dos GDs. Em Liang e Wang (2003), é proposto um método de controle coordenado de módulos de BCs e dos taps dos RTs. Embora os BCs, RTs e GDs tenham alta presença nos sistemas de distribuição atuais, o ajuste ótimo das correspondentes variáveis de controle não tem sido completamente estudado. Aliás, o controle coordenado considerando todos estes elementos tem sido reportado em poucos trabalhos, dentre os quais destacam-se Gonçalves et al. (2013) e Araújo, Meira e Almeida (2013).

Uma característica comum nos trabalhos citados no parágrafo anterior é a necessidade de conhecer de forma detalhada o modelo físico do SDEE, isto é, a necessidade do conhecimento completo dos parâmetros elétricos e do perfil de demanda do SDEE. Um problema comum é que normalmente essas informações estão desatualizadas, não existem, não podem ser facilmente estimadas, ou não são confiáveis.

Na tendência das *smart grids*, os modernos SDEE estão migrando de uma estrutura não supervisionada e passiva para uma estrutura supervisionada e ativa devido à instalação de dispositivos de medição e à proliferação do uso de sistemas digitais de proteção (MOMOH, 2017), (NORTHCOTE-GREEN; WILSON, 2006). Neste contexto, é inevitável o desenvolvimento de novas técnicas de controle, monitoramento, automação e proteção específicas para esses modernos SDEE. Uma característica importante que essas novas técnicas devem incorporar refere-se

a como usar as informações relevantes dos sistemas de monitoramento e medição que estão disponíveis em diversos pontos da rede elétrica, as quais podem ser coletadas e armazenadas para seu uso nas tomadas de decisões de forma centralizada e otimizada. Um importante benefício que pode ser obtido dessa estrutura é o controle centralizado ótimo da magnitude de tensão e da potência reativa no SDEE (JAHANGIRI; ALIPRANTIS, 2013; MOHAPATRA; BIJWE; PANIGRAHI, 2014; NIKNAM; ZARE; AGHAEI, 2012; ROYTELMAN; GANESAN, 2000).

Neste contexto, os sistemas inteligentes (SIs) podem ser aplicados com o intuito de gerenciar um sistema de controle que permita operar o SDEE adequadamente.

Os SIs definem o ajuste mais apropiado da geração de potencia ativa e reativa dos GDs, o número de módulos em operação dos BCs, e a posição do tap para os OLTC e RTs, com o objetivo de minimizar as perdas de potencia ativa do sistema e controlar as tensões usando as medições disponíveis no centro de controle. Os SIs podem ser treinados a partir das medições disponíveis e ações registradas no centro de controle do sistema. Entretanto, é usado um modelo matemático de programação linear inteiro misto (MPLIM) proposto em Gonçalves et al. (2013) para simular este conjunto de medições e ações. O software WEKA, versão 3.8.0, foi utilizado para gerenciar os três SIs utilizados. WEKA usa os algoritmos de aprendizado *Multilayer Perceptron/Backpropagation*, o J48/C4.5 e o SMO/PoliyKernel para realizar o treinamento das RNAs, ADs e MVSs, respetivamente. O modelo MPLIM foi implementado na linguagem de modelagem matemática AMPL (FOURER; GAY; KERNIGHAN, 2002) e foi resolvido usando o solver comercial CPLEX (CPLEX, 2009). Para demonstrar a eficiência da técnica de solução proposta, vários testes foram realizados usando um sistema de distribuição de 42 nós.

3.1.1 Formulação Cônica de Segunda Ordem Binário Misto para o Problema de POSD Radiais

O problema de planejamento da operação dos sistemas de distribuição (POSD) é um problema de programação não linear inteiro misto(PNLIM) de difícil solução que pode ser representado matematicamente como um modelo de programação cônica de segunda ordem binário misto, (PCSOBM). No modelo apresentado é considerada a existência de GDs, RTs e BCs. A seguir, apresenta-se a formulação cônica de segunda ordem binário misto para o problema de POSD.

Com este modelo determinaram-se as injeções de potência ativa e reativa dos GDs, o número de módulos de capacitores em operação e o número de passos do tap dos RTs, com a finalidade de minimizar o custo das perdas diárias de energia do sistema, representadas na função objetivo ($\min v$) do POSD.

$$\min v = \sum_{d\in\Omega_d} c_d^{ls}\alpha_d \sum_{ij\in\Omega_l} R_{ij} I_{ij,d}^{qdr}$$

sujeito a:

$$\sum_{ki\in\Omega_l} P_{ki,d} - \sum_{ij\in\Omega_l}(P_{ij,d} + R_{ij}I_{ij,d}^{qdr}) + \sum_{ki\in\Omega_r t} P_{ki,d}^{rt} - \sum_{ij\in\Omega_r t} P_{ij,d}^{rt} + P_{i,d}^{s} + \sum_{\substack{m\in\Omega_{gd}/\\ i=L_{gd}(m)}} P_{m,d}^{gd} = P_{i,d}^{D} \qquad \forall i\in\Omega_b, \forall t\in\Omega_t \tag{2}$$

$$\sum_{ki\in\Omega_l} Q_{ki,d} - \sum_{ij\in\Omega_l}(Q_{ij,d} + X_{ij}I_{ij,d}^{qdr}) + \sum_{ki\in\Omega_{rt}} Q_{ki,d}^{rt} - \sum_{ij\in\Omega_r t} Q_{ij,d}^{rt} + Q_{i,d}^{S} + \sum_{\substack{m\in\Omega_{gd}/\\ i=L_{gd}(m)}} Q_{m,d}^{gd} \sum_{\substack{n\in\Omega_{bc}/\\ i=L_{bc}(m)}} Q_{n,d}^{bc} = Q_{i,d}^{D} \qquad \forall i\in\Omega_b, \forall d\in\Omega_d \tag{3}$$

$$V_{i,d}^{qdr} - 2(R_{ij}P_{ij,d} + X_{ij}Q_{ij,d}) - Z_{ij}^2 I_{ij,d}^{qdr} - V_{j,d}^{qdr} = 0 \qquad \forall ij\in\Omega_l, \forall d\in\Omega_d \tag{4}$$

$$I_{ij,d}^{qdr} V_{j,d}^{qdr} \geq P_{ij,d}^2 + Q_{ij,d}^2 \qquad \forall ij\in\Omega_l, \forall d\in\Omega_d \tag{5}$$

$$(P_{m,d}^{gd})^2 + (Q_{m,d}^{gd})^2 \leq (S_m^{-gd})^2 \qquad \forall m\in\Omega_{dg}, \forall d\in\Omega_d \tag{6}$$

$$0 \leq P_{m,d}^{gd} \qquad \forall m\in\Omega_{dg}, \forall d\in\Omega_d \tag{7}$$

$$-P_{m,d}^{gd}\tan(\cos^{-1}(\underline{fp}_m^{gd})) \leq Q_{m,d}^{gd} \leq P_{m,d}^{gd}$$

$$\tan(\cos^{-1}(\overline{fp}_m^{gd})) \qquad \forall m\in\Omega_{dg}, \forall d\in\Omega_d \tag{8}$$

$$Q_{n,d}^{bc} = na_{n,d}^{bc} Q_n^{esp} \qquad \forall n\in\Omega_{bc}, \forall d\in\Omega_d \tag{9}$$

$$|na_{n,d}^{bc} - na_{n,d-1}^{bc}| \leq \triangle_n^{-bc} \qquad \forall n\in\Omega_{bc}, \forall d\in\Omega_d | d>1 \tag{10}$$

$$0 \leq na_{n,d}^{bc} \leq \overline{na}_n^{bc} \qquad \forall n\in\Omega_{bc}, \forall d\in\Omega_d \tag{11}$$

$$V_{j,d}^{qdr} = \sum_{k=0\ldots 2\overline{nt}_{ij}} [(1 + R_{i,j}^{\%}\frac{(k-\overline{nt}_{i,j})}{\overline{nt}_{i,j}})^2 V_{i,j,d,k}^{c}]$$

$$\forall ij\in\Omega_{rt}, \forall d\in\Omega_d \tag{12}$$

$$\sum_{k=0\ldots 2\overline{nt}_{ij}} bt_{ij,d,k} = 1 \qquad \forall i,j\in\Omega_{rt}, \forall d\in\Omega_d \tag{13}$$

$$(\underline{V^2})bt_{ij,d,k} \leq V_{i,j,d,k}^{c} \leq (\overline{V^2})bt_{ij,d,k}$$

$$\forall ij\in\Omega_{rt}, \forall d\in\Omega_d, \forall k = 0\ldots 2\overline{nt}_{ij} \tag{14}$$

$$(\underline{V^2})(1 - bt_{ij,d,k}) \leq V_{i,d}^{qdr} - V_{i,j,d,k}^{c} \leq (\overline{V^2})(1 - bt_{ij,d,k})$$

$$\forall ij\in\Omega_{rt}, \forall d\in\Omega_d, \forall k = 0\ldots 2\overline{nt}_c \tag{15}$$

$$|\sum_{k=0\ldots 2\overline{nt}_{ij}} [(k-\overline{nt}_{ij})bt_{ij,d,k}] - \sum_{k=0\ldots 2\overline{nt}_{ij}} [(k-\overline{nt}_{ij})$$

$$bt_{ij,d-1,k}]| \leq \overline{\triangle}_{i,j}^{rt} \qquad \forall i,j\in\Omega_{rt}, \forall d\in\Omega_d | d>1 \tag{16}$$

$$(\underline{V^2}) \leq V_{i,d}^{qdr} \leq (\overline{V^2}) \qquad \forall i \in \Omega_b, \forall d \in \Omega_d \tag{17}$$

$$0 \leq I_{ij,d}^{qdr} \leq \overline{I}_{ij}^2 \qquad \forall ij \in \Omega_l, \forall d \in \Omega_d \tag{18}$$

$$na_{n,d}^{bc} \quad inteiro \qquad \forall n \in \Omega_{bc}, \forall d \in \Omega_d \tag{19}$$

$$bt_{ij,d,k} \quad binario \qquad \forall ij \in \Omega_{rt}, \forall d \in \Omega_d, \forall k = 0...2\overline{nt}_{ij} \tag{20}$$

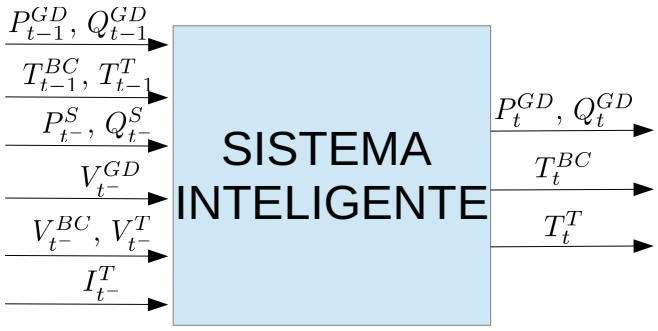
A descrição das restrições deste problema são apresentadas a seguir.

As restrições (2) e (3) representam o balanço de potência ativa e reativa, respetivamente. Em (2) considera-se a presença de GDs e em (3) considera-se a presença de GDs e BCs. As equações (4) e (5) representam, respetivamente, o cálculo da queda de tensão entre dois nós consecutivos e a corrente nos circuitos. Em (6) é representada a capacidade máxima de geração de potência aparente dos GDs. Em (7) e (8) é representada a potência reativa e o limite do fator de potência dos GDs, respetivamente, nos níveis de demanda. Em (9) representa-se a capacidade de potência reativa dos módulos dos BCs chaveados. A restrição (10) representa a máxima variação do número de módulos de capacitores em operação em horas consecutivas. Em (11) são representados o número de módulos de capacitores instalados nos BCs chaveados. O cálculo da tensão regulada é definido por (12). A transformação em variáveis binárias do passo dos taps dos RTs é definida por (13). Várias das restrições não lineares relacionadas com a tensão são linearizadas e representadas pelas restrições lineares (14) e (15) (formulação disjuntiva). A restrição (16) representa o cálculo da máxima variação de passo dos taps dos RTs em horas consecutivas. Em (17) se representam os limites do quadrado da magnitude de tensão nos nós do sistema, enquanto que (18) representa os limites do quadrado da magnitude do fluxo de corrente nos circuitos. As restrições (9)-(11) e (19) representam o modelo matemático do BC chaveado. As restrições (19) e (20) determinam a natureza das variáveis, ou seja, se são variáveis inteiras ou binárias.

3.2 METODOLOGIA PROPOSTA

Nesta tese é apresentada uma metodologia baseada em SIs, que permite determinar o controle centralizado de magnitude de tensão e potencia reativa (Volt-VAr) em tempo real dos SDEE, usando medições elétricas como base de dados de entrada, estes dados são submetidos a um pré-processamento para posteriormente ingressá-los ao sistema inteligente. Dentro do SI é realizado o processo de classificação e finalmente define-se o ajuste mais apropriado da geração de potência ativa e reativa dos GDs, o número de módulos em operação dos BCs, e a posição do tap para os transformadores com comutação de tap sob carga e RTs. Na Figura 16, apresenta-se um diagrama que mostra as respetivas variáveis de entradas e saídas do SI.

Figura 16 - Diagrama das variáveis de entradas e saídas do SI.



Fonte: Próprio Autor

Para descrever o funcionamento dos SIs, é necessário conhecer os dados de entrada e de saída. Como dados de entrada têm-se:

- P^{GD}_{t-1}: Valores da injeção de potências ativa dos geradores 1 e 2 na hora $t-1$;
- Q^{GD}_{t-1}: Valores da injeção de potência reativa do gerador 1 e 2 na hora $t-1$;
- T^{BC}_{t-1}: Valores dos módulos em operação dos bancos de capacitores 1, 2, 3, 4 e 5 na hora $t-1$;
- T^{T}_{t-1}: Valores do número de passos do tap dos transformadores com comutação de tap sob carga e regulador de tensão 1, 2, 3 e 4 na hora $t-1$;
- $P^{S}_{t^-}$: Injeção de potência ativa da subestação na hora t^-;
- $Q^{S}_{t^-}$: Injeção de potência reativa da subestação na hora t^-;
- $V^{GD}_{t^-}$: Magnitude de tensão nos nós dos geradores distribuídos 1 e 2 em operação na hora t^-;
- $V^{BC}_{t^-}$: Magnitude de tensão nos nós dos bancos de capacitores 1, 2, 3, 4 e 5 em operação na hora t^-;
- $V^{T}_{t^-}$: Magnitude de tensão nos nós controlados pelos transformadores com comutação de tap sob carga e regulador de tensão 1, 2, 3 e 4 na hora t^-;
- $I^{T}_{t^-}$: Magnitude de corrente nos transformadores com comutação de tap sob carga e regulador de tensão 1, 2, 3 e 4 na hora t^-;

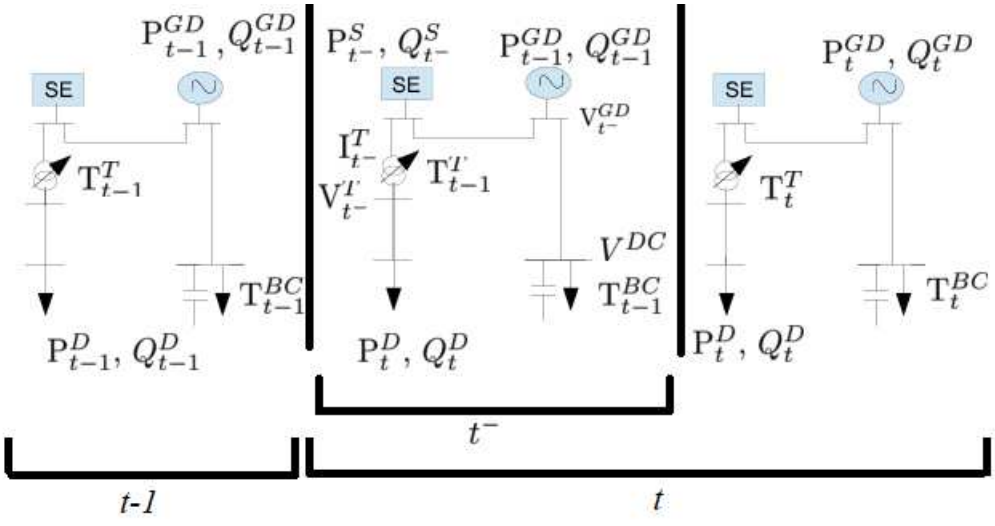
Como dados de saída dos SIs, têm-se as ações de controle na hora t:

- P^{GD}_{t}: Valores da injeção de potência ativa dos geradores 1 e 2 na hora t;

- Q_t^{GD}: Valores da injeção de potência reativa dos geradores 1 e 2 na hora t;
- T_t^{BC}: Valores dos módulos em operação dos bancos de capacitores 1, 2, 3, 4 e 5 na hora t;
- T_t^T: Valores do número de passos do tap dos transformadores com comutação de tap sob carga e regulador de tensão 1, 2, 3 e 4 na hora t;

Para entender a geração dos dados de entrada e de saída nos SIs, uma representação gráfica dos passos necessários na criação da base de dados utilizada para realizar o treinamento dos SIs é mostrado na Fig. 17. Considere que no período de tempo $t-1$, em que o sistema está operando em regime permanente com uma demanda P_{t-1}^D e Q_{t-1}^D e com os valores das variáveis de controle P_{t-1}^{GD}, Q_{t-1}^{GD}, T_{t-1}^{BC}, T_{t-1}^T. Note que entre o tempo $t-1$ e t há um tempo que é chamado de $t-$, este tempo é aquele que está antes das mudanças do controle do sistema, isto quer dizer que os controles definidos no periodo de tempo $t-1$ continuam fixos em quanto a demanda varia. Dados os controles definidos no periodo de tempo $t-1$ fixos e uma variação da demanda (P_t^D, Q_t^D), um novo ponto de operação é definido no sistema no tempo t^-. Para este caso de estudos, as medidas são as representadas pelas variáveis de controle: $P_{t^-}^S$, $Q_{t^-}^S$, $V_{t^-}^{GD}$, $V_{t^-}^{BC}$, $V_{t^-}^T$ e $I_{t^-}^T$. Assim é possivel obter todos os dados de entrada dos SIs. No período de transição entre o tempo t^- e t, a demanda do sistema continua fixa, isto é, P_t^D e Q_t^D. As variáveis de controle P_t^{GD}, Q_t^{GD}, T_t^{BC} e T_t^T no tempo t são calculadas usando um modelo MPLIM apresentado em (GONÇALVES et al., 2013), que representam os dados de saída dos SIs. Em (VILLACCI; BONTEMPI; VACCARO, 2006) foi utilizada uma metodologia similar à usada neste trabalho, para criação da base de dados.

Figura 17 - Descrição das variáveis de controle nos diferentes instantes de tempo.



Fonte: Próprio Autor

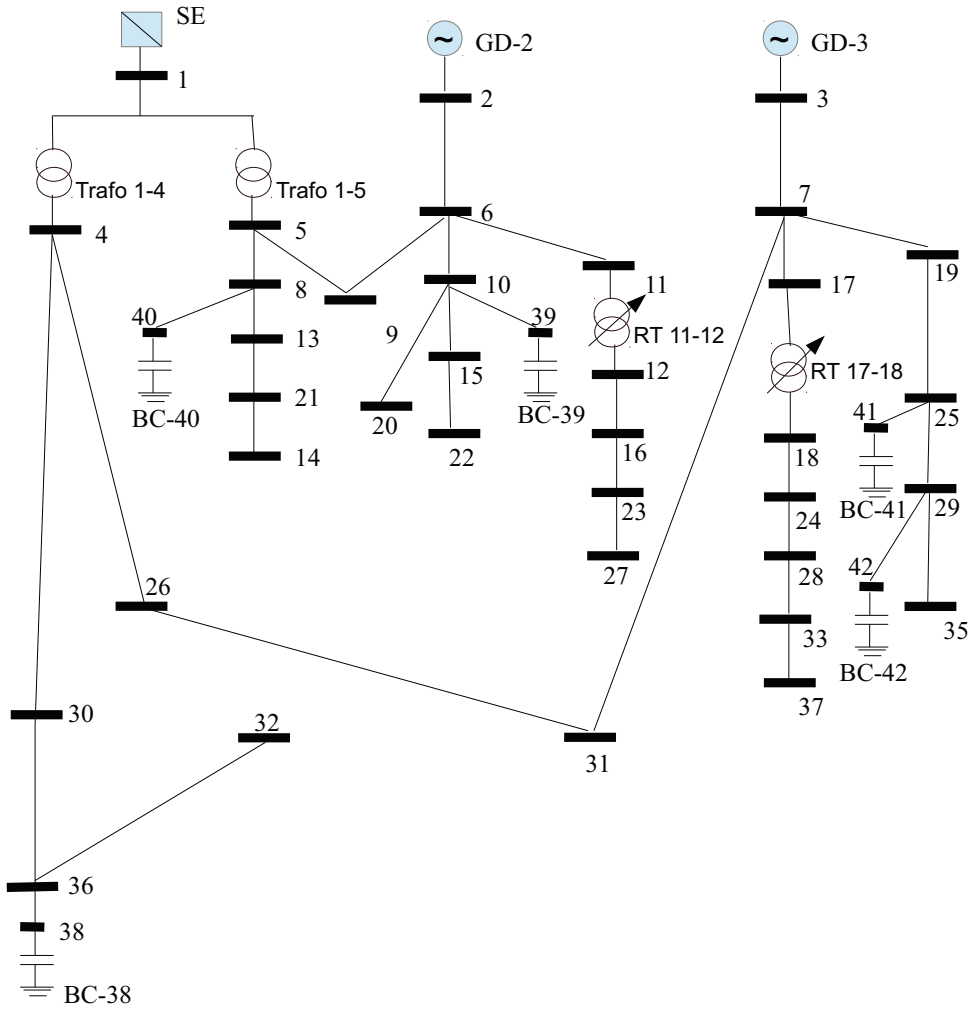
Para realizar o treinamento dos SI é necessário criar bancos de dados com todas as medições do sistema no tempo $t-1$, no tempo $t-$ e as ações de controle no tempo t. Neste trabalho foi

realizada uma análise do sistema durante 365 dias, hora a hora, portanto foram obtidos em total 8760 cenários. Estes cenários foram usados para realizar o treinamento dos SIs. Uma vez tendo o sistema treinado, foi criado outro banco de dados com cenários de 7 dias, hora a hora, para realizar a avaliação dos SIs. Para a aplicação desta metodologia foi utilizada a plataforma de aprendizado de máquina WEKA.

3.3 CASO DE ESTUDO E RESULTADOS

O diagrama unifilar ilustrado na Figura 18, representa a configuração do sistema que foi usado em (GONÇALVES et al., 2013) e que também foi utilizado para realizar o desenvolvimento do caso de estudo apresentado neste capítulo.

Figura 18 - Sistema de distribuição de 42 nós.



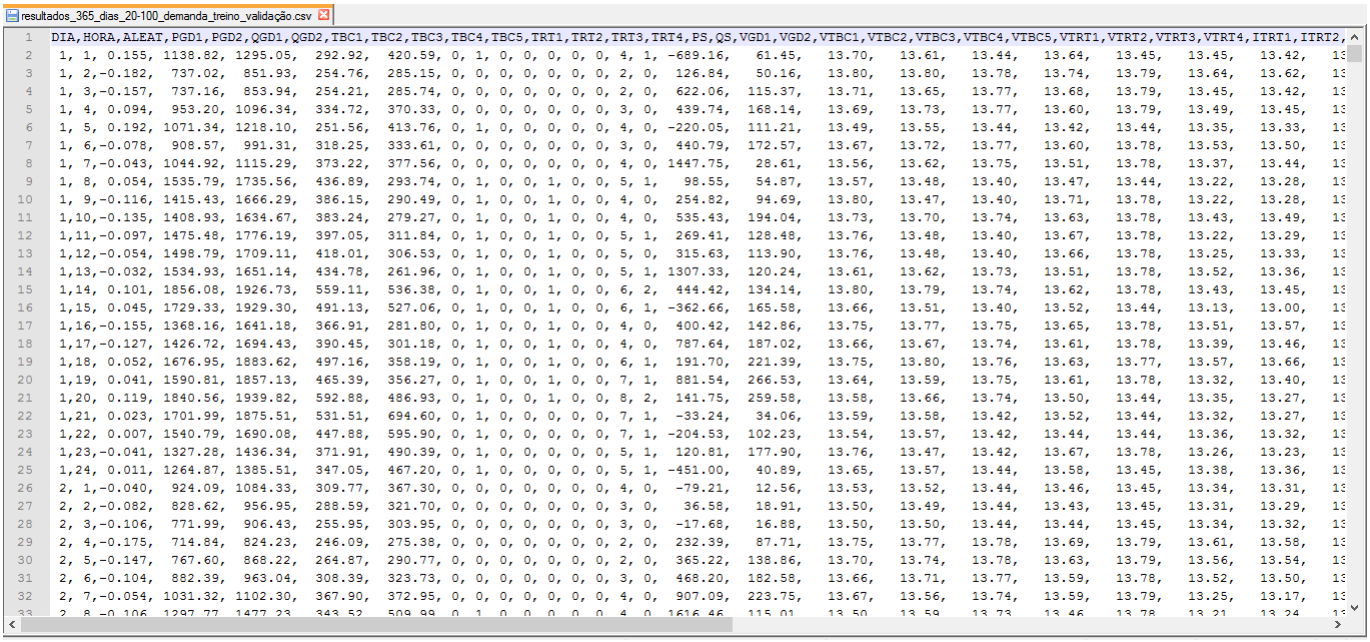
Fonte: Adaptado de Gonçalves et al. (2013)

Com o objetivo de treinar e validar a metodologia proposta foram realizados testes com

diferentes percentagens de variação aleatória da demanda (5%, 10%, 15% e 20%). Segundo os testes realizados com os SIs, os dados de treinamento apresentam um alto nível de veracidade na sua validação. Isto pode ser conferido nas figuras e tabelas com os resultados obtidos que serão apresentadas nesta seção.

Na Figura 19 se ilustra uma das bancos de dados que foram simuladas com o uso do modelo MPLIM. Estes dados são utilizados para realizar o treinamento e validação dos SIs. Este banco de dados, em particular, está composto por medidas hora a hora de 33 atributos, por um período de 365 dias, portanto, estão sendo analisados inicialmente 289080 dados. Cabe salientar que foram utilizadas diferentes bancos de dados tanto no processo de treinamento quanto no processo de validação.

Figura 19 - Banco de dados de treinamento.



resultados_365_dias_20-100_demanda_treino_validação.csv

```
DIA,HORA,ALEAT,PGD1,PGD2,QGD1,QGD2,TBC1,TBC2,TBC3,TBC4,TBC5,TRT1,TRT2,TRT3,TRT4,PS,QS,VGD1,VGD2,VTBC1,VTBC2,VTBC3,VTBC4,VTBC5,VTRT1,VTRT2,VTRT3,VTRT4,ITRT1,ITRT2,
 1, 1, 0.155, 1138.82, 1295.05,  292.92,  420.59, 0, 1, 0, 0, 0, 0, 4, 1, -689.16,   61.45,  13.70,  13.61,  13.44,  13.64,  13.45,  13.45,  13.42,  13
 1, 2,-0.182,  737.02,  851.93,  254.76,  285.15, 0, 0, 0, 0, 0, 0, 2, 0,  126.84,   50.16,  13.80,  13.80,  13.78,  13.74,  13.79,  13.64,  13.62,  13
 1, 3,-0.157,  737.16,  853.94,  254.21,  285.74, 0, 0, 0, 0, 0, 0, 2, 0,  622.06,  115.37,  13.71,  13.65,  13.77,  13.68,  13.79,  13.45,  13.42,  13
 1, 4, 0.094,  953.20, 1096.34,  334.72,  370.33, 0, 0, 0, 0, 0, 0, 3, 0,  439.74,  168.14,  13.69,  13.73,  13.77,  13.60,  13.79,  13.49,  13.45,  13
 1, 5, 0.192, 1071.34, 1218.10,  251.56,  413.76, 0, 1, 0, 0, 0, 0, 4, 0, -220.05,  111.21,  13.49,  13.55,  13.44,  13.42,  13.44,  13.35,  13.33,  13
 1, 6,-0.078,  908.57,  991.31,  318.25,  333.61, 0, 0, 0, 0, 0, 0, 3, 0,  440.79,  172.57,  13.67,  13.72,  13.77,  13.60,  13.78,  13.53,  13.50,  13
 1, 7,-0.043, 1044.92, 1115.29,  373.22,  377.56, 0, 0, 0, 0, 0, 0, 4, 0, 1447.75,   28.61,  13.56,  13.62,  13.75,  13.51,  13.78,  13.37,  13.44,  13
 1, 8, 0.054, 1535.79, 1735.56,  436.89,  293.74, 0, 1, 0, 0, 1, 0, 0, 5, 1,   98.55,   54.87,  13.57,  13.48,  13.40,  13.47,  13.44,  13.22,  13.28,  13
 1, 9,-0.116, 1415.43, 1666.29,  386.15,  290.49, 0, 1, 0, 0, 1, 0, 0, 4, 0,  254.82,   94.69,  13.80,  13.47,  13.40,  13.71,  13.78,  13.22,  13.28,  13
 1,10,-0.135, 1408.93, 1634.67,  383.24,  279.27, 0, 1, 0, 0, 1, 0, 0, 4, 0,  535.43,  194.04,  13.73,  13.70,  13.74,  13.63,  13.78,  13.43,  13.49,  13
 1,11,-0.097, 1475.48, 1776.19,  397.05,  311.84, 0, 1, 0, 0, 1, 0, 0, 5, 1,  269.41,  128.48,  13.76,  13.48,  13.40,  13.67,  13.78,  13.22,  13.29,  13
 1,12,-0.054, 1498.79, 1709.11,  418.01,  306.53, 0, 1, 0, 0, 1, 0, 0, 5, 0,  315.63,  113.90,  13.76,  13.48,  13.40,  13.66,  13.78,  13.25,  13.33,  13
 1,13,-0.032, 1534.93, 1651.14,  434.78,  261.96, 0, 1, 0, 0, 1, 0, 0, 5, 1, 1307.33,  120.24,  13.61,  13.62,  13.73,  13.51,  13.78,  13.52,  13.36,  13
 1,14, 0.101, 1856.08, 1926.73,  559.11,  536.38, 0, 1, 0, 0, 1, 0, 0, 6, 2,  444.42,  134.14,  13.80,  13.79,  13.74,  13.62,  13.78,  13.43,  13.45,  13
 1,15, 0.045, 1729.33, 1929.30,  491.13,  527.06, 0, 1, 0, 0, 1, 0, 0, 6, 1, -362.66,  165.58,  13.66,  13.51,  13.40,  13.52,  13.44,  13.13,  13.00,  13
 1,16,-0.155, 1368.16, 1641.18,  366.91,  281.80, 0, 1, 0, 0, 1, 0, 0, 4, 0,  400.42,  142.86,  13.75,  13.77,  13.75,  13.65,  13.78,  13.51,  13.57,  13
 1,17,-0.127, 1426.72, 1694.43,  390.45,  301.18, 0, 1, 0, 0, 1, 0, 0, 4, 0,  787.64,  187.02,  13.66,  13.67,  13.74,  13.61,  13.78,  13.39,  13.46,  13
 1,18, 0.052, 1676.95, 1883.62,  497.16,  358.19, 0, 1, 0, 0, 1, 0, 0, 6, 1,  191.70,  221.39,  13.75,  13.80,  13.76,  13.63,  13.77,  13.57,  13.66,  13
 1,19, 0.041, 1590.81, 1857.13,  465.39,  356.27, 0, 1, 0, 0, 1, 0, 0, 7, 1,  881.54,  266.53,  13.64,  13.59,  13.75,  13.61,  13.78,  13.32,  13.40,  13
 1,20, 0.119, 1840.56, 1939.82,  592.88,  486.93, 0, 1, 0, 0, 1, 0, 0, 8, 2,  141.75,  259.58,  13.58,  13.66,  13.74,  13.50,  13.44,  13.35,  13.27,  13
 1,21, 0.023, 1701.99, 1875.51,  531.51,  694.60, 0, 1, 0, 0, 0, 0, 0, 7, 1,  -33.24,   34.06,  13.59,  13.58,  13.42,  13.52,  13.44,  13.32,  13.27,  13
 1,22, 0.007, 1540.79, 1690.08,  447.88,  595.90, 0, 1, 0, 0, 0, 0, 0, 7, 1, -204.53,  102.23,  13.54,  13.57,  13.42,  13.44,  13.44,  13.36,  13.32,  13
 1,23,-0.041, 1327.28, 1436.34,  371.91,  490.39, 0, 1, 0, 0, 0, 0, 0, 5, 1,  120.81,  177.90,  13.76,  13.47,  13.42,  13.67,  13.78,  13.26,  13.23,  13
 1,24, 0.011, 1264.87, 1385.51,  347.05,  467.20, 0, 1, 0, 0, 0, 0, 0, 5, 1, -451.00,   40.89,  13.65,  13.57,  13.44,  13.58,  13.45,  13.38,  13.36,  13
 2, 1,-0.040,  924.09, 1084.33,  309.77,  367.30, 0, 0, 0, 0, 0, 0, 0, 4, 0,  -79.21,   12.56,  13.53,  13.52,  13.44,  13.46,  13.45,  13.34,  13.31,  13
 2, 2,-0.082,  828.62,  956.95,  288.59,  321.70, 0, 0, 0, 0, 0, 0, 0, 3, 0,   36.58,   18.91,  13.50,  13.49,  13.44,  13.43,  13.45,  13.31,  13.29,  13
 2, 3,-0.106,  771.99,  906.43,  255.95,  303.95, 0, 0, 0, 0, 0, 0, 0, 3, 0,  -17.68,   16.88,  13.50,  13.50,  13.44,  13.44,  13.45,  13.34,  13.32,  13
 2, 4,-0.175,  714.84,  824.23,  246.09,  275.38, 0, 0, 0, 0, 0, 0, 0, 2, 0,  232.39,   87.71,  13.75,  13.77,  13.78,  13.69,  13.79,  13.61,  13.58,  13
 2, 5,-0.147,  767.60,  868.22,  264.87,  290.77, 0, 0, 0, 0, 0, 0, 0, 2, 0,  365.22,  138.86,  13.70,  13.74,  13.78,  13.63,  13.79,  13.56,  13.54,  13
 2, 6,-0.104,  882.39,  963.04,  308.39,  323.73, 0, 0, 0, 0, 0, 0, 0, 3, 0,  468.20,  182.58,  13.66,  13.71,  13.77,  13.59,  13.78,  13.52,  13.50,  13
 2, 7,-0.054, 1031.32, 1102.30,  367.90,  372.95, 0, 0, 0, 0, 0, 0, 0, 4, 0,  907.09,  223.75,  13.67,  13.56,  13.74,  13.59,  13.79,  13.25,  13.17,  13
 2, 8,-0.106, 1297.77, 1477.23,  343.52,  509.99, 0, 1, 0, 0, 0, 0, 0, 4, 0, 1616.46,  115.01,  13.50,  13.59,  13.73,  13.46,  13.78,  13.21,  13.24,  13
```

Fonte: Próprio Autor

Os atributos que foram escolhidos para realizar o treinamento dos SIs são:

- DIA;
- HORA;
- ALEAT: Valor de variação aleatória da demanda;
- PGD1: Valor da injeção de potência ativa do gerador 1;
- PGD2: Valor da injeção de potência ativa do gerador 2;

- QGD1: Valor da injeção de potência reativa do gerador 1;
- QGD2: Valor da injeção de potência reativa do gerador 2;
- TBC1: Valor do módulo em operação do banco de capacitores 1;
- TBC2: Valor do módulo em operação do banco de capacitores 2;
- TBC3: Valor do módulo em operação do banco de capacitores 3;
- TBC4: Valor do módulo em operação do banco de capacitores 4;
- TBC5: Valor do módulo em operação do banco de capacitores 5;
- TRT1: Valor do número de passos do tap dos transformadores com comutação de tap sob carga e regulador de tensão 1;
- TRT2: Valor do número de passos do tap dos transformadores com comutação de tap sob carga e regulador de tensão 2;
- TRT3: Valor do número de passos do tap dos transformadores com comutação de tap sob carga e regulador de tensão 3;
- TRT4: Valor do número de passos do tap dos transformadores com comutação de tap sob carga e regulador de tensão 4;
- PS: Injeção de potência ativa da subestação;
- QS: Injeção de potência reativa da subestação;
- VGD1: Magnitude de tensão nos nós do gerador distribuído 1 em operação;
- VGD2: Magnitude de tensão nos nós do gerador distribuído 2 em operação;
- VTBC1: Magnitude de tensão nos nós do banco de capacitores 1 em operação;
- VTBC2: Magnitude de tensão nos nós do banco de capacitores 2 em operação;
- VTBC3: Magnitude de tensão nos nós do banco de capacitores 3 em operação;
- VTBC4: Magnitude de tensão nos nós do banco de capacitores 4 em operação;
- VTBC5: Magnitude de tensão nos nós do banco de capacitores 5 em operação;
- VTRT1: Magnitude de tensão nos nós controlados pelos transformadores com comutação de tap sob carga e regulador de tensão 1;
- VTRT2: Magnitude de tensão nos nós controlados pelos transformadores com comutação de tap sob carga e regulador de tensão 2;

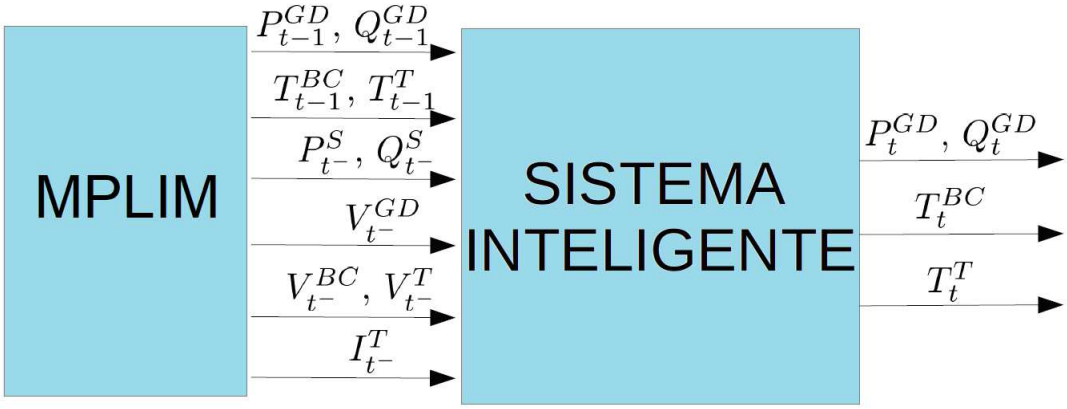
- VTRT3: Magnitude de tensão nos nós controlados pelos transformadores com comutação de tap sob carga e regulador de tensão 3;
- VTRT4: Magnitude de tensão nos nós controlados pelos transformadores com comutação de tap sob carga e regulador de tensão 4;
- ITRT1: Magnitude de corrente nos transformadores com comutação de tap sob carga e regulador de tensão 1;
- ITRT2: Magnitude de corrente nos transformadores com comutação de tap sob carga e regulador de tensão 2;
- ITRT3: Magnitude de corrente nos transformadores com comutação de tap sob carga e regulador de tensão 3;
- ITRT4: Magnitude de corrente nos transformadores com comutação de tap sob carga e regulador de tensão 4;
- output: Controle definido para o sistema na hora t;

Para uma melhor compreensão de cada processo que compõe as diferentes etapas de aprendizado dos SIs, são apresentados o processo de treinamento e de validação separadamente.

3.3.1 Treinamento dos SIs

Os SIs são treinados e avaliados separadamente usando diferentes bancos de dados. A forma de treinar os SIs neste caso de estudo é realizada da seguinte maneira: inicialmente MPLIM entrega os valores simulados para cada uma das variáveis de controle. Estes valores são repassados para o SI quem se encarrega de definir o novo controle do sistema a partir dos dados recebidos. Posteriormente, o resultado entregue pelo SI é repassado novamente para o MPLIM quem vai avaliar a qualidade do resultado comparando a tensão e as perdas geradas a partir da utilização do controle definido pelo SI. Este processo é repetido hora pós hora durante os 365 dias do ano. Na Figura 20 é ilustrado o processo realizado no treinamento do SI. Com esta forma de treinamento, pretende-se ensinar aos SIs a decidir que ações realizar a partir de determinado comportamento do sistema.

Figura 20 - Processo de treinamento realizado entre o MPLIM e o SI.



Fonte: Próprio Autor

Para analisar o grau de aprendizado dos SIs durante o processo de treinamento, foram criadas quatro bancos de dados com o histórico de perdas de energia do sistema a cada hora durante uma semana, para avaliar cada um dos SIs utilizados. Cada banco de dados continha dados criados a partir de uma variação aleatória de demanda de 5%, 10%, 15% e 20% respetivamente.

Na Tabela 1, são comparados os resultados obtidos pelos SIs e os resultados obtidos após a simulação com o modelo MPLIM. Considerando que a variação aleatória de demanda de 20% pode representar um maior esforço computacional devido que o SI terá que aprender a realizar o melhor controle a partir de uma variação deste tamanho, compara-se o desempenho de cada SI proposto para este caso em particular.

Tabela 1 - Perdas de energia do sistema obtidas durante uma semana usando as diferentes metodologias

Perdas de energia em kWh (20% alteração da demanda)							
Metodologia	**Dia 1**	**Dia 2**	**Dia 3**	**Dia 4**	**Dia 5**	**Dia 6**	**Dia 7**
MPLIM	11352,61	10202,87	10139,05	10441,28	9464,15	9387,92	10266,44
MVS	11888,01	10684,05	10617,22	10933,70	9910,49	9830,67	10750,61
AD	12090,73	10866,24	10798,27	11120,15	10079,49	9998,31	10933,94
RNA	12288,27	11043,77	10974,69	11301,83	10244,17	10161,66	11112,57

Fonte: Dados da pesquisa do autor.

Assim, para avaliar os resultados do treinamento dos SIs, escolhe-se desta base de dados, o dia 1, que é o dia que apresenta maiores níveis de perdas de energia. Na Tabela 2 é realizada uma comparação dos valores hora a hora das perdas de energia do sistema em estudo, calculados pelo MPLIM e definidos pelos três algoritmos que compõem os SIs, para o dia 1 que é o dia que apresenta maiores níveis de demanda e maiores níveis de perdas de energia. Cabe salientar que para realizar o cálculo das perdas de energia a partir do controle definido pelos SIs, é utilizado um fluxo de carga dado que os SI não realizam calculo algum.

Na Figura 21, apresentam-se as perdas de potência no sistema para cada intervalo de tempo

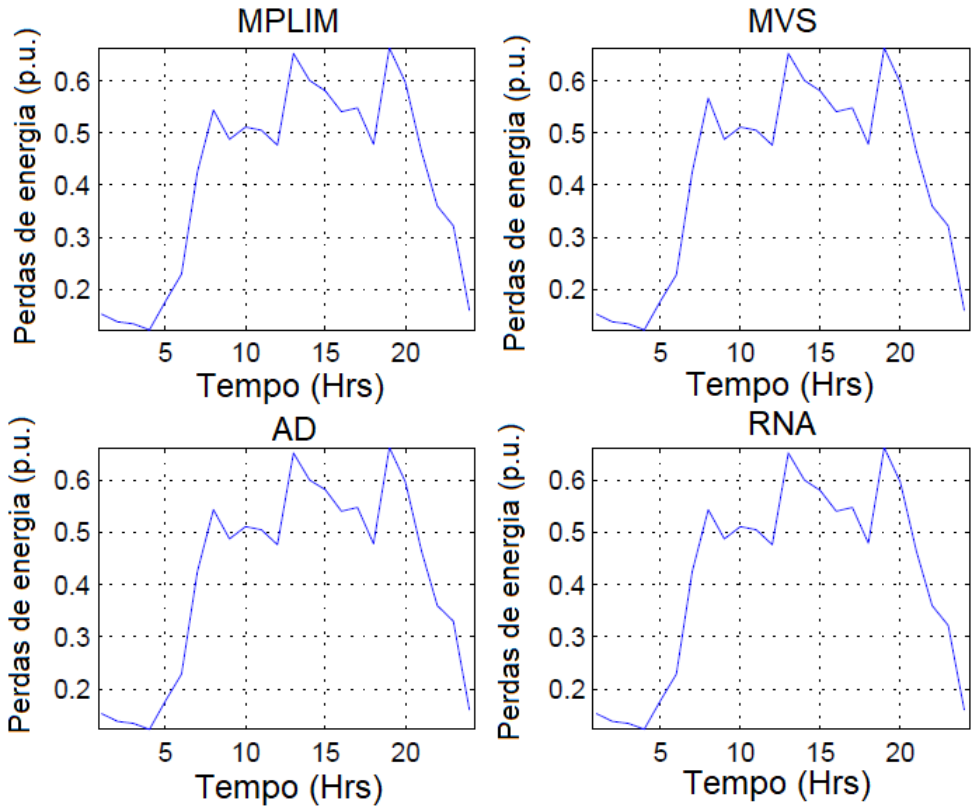
Tabela 2 - Valores calculados pelo MPLIM e definidos pelo SI das perdas de energia do sistema hora a hora em kWh

	MPLIM	**MVS**	**AD**	**RNA**
Hora	**Dia 1**	**Dia 1**	**Dia 1**	**Dia 1**
0	63,92	69,90	71,95	71,57
1	71,78	71,95	71,78	71,78
2	63,92	71,78	72,15	71,78
3	71,78	71,95	72,15	83,13
4	71,95	71,95	72,15	83,13
5	83,13	87,54	91,40	91,40
6	562,01	586,73	586,73	586,73
7	531,70	562,01	586,73	586,73
8	1121,24	1146,07	1146,07	1159,75
9	586,73	586,73	586,73	655,90
10	562,01	586,73	586,73	586,73
11	562,02	586,77	586,73	586,73
12	531,70	586,77	586,73	562,01
13	586,73	586,77	586,77	655,90
14	531,70	586,73	562,01	562,01
15	562,01	586,73	586,77	562,01
16	880,87	880,87	905,62	905,62
17	507,00	531,70	531,70	586,73
18	1146,07	1159,75	1159,75	1159,75
19	1146,07	1146,07	1159,75	1159,75
20	507,00	531,70	586,73	586,73
21	343,27	383,18	485,96	485,96
22	189,56	335,69	335,69	343,26
23	68,46	71,95	71,95	83,13
Total	**11352,61**	**11888,01**	**12090,73**	**12288,22**

Fonte: Dados da pesquisa do autor.

(ação de controle definida pelos algoritmos que compõem os SIs durante 24 horas) para a percentagem de variação de demanda de 20%. Nesta figura, os valores das perdas obtidos usando a ação de controle ótima obtidos na simulação do modelo MPLIM, e os outros, correspondem à ação de controle definida pelos SIs.

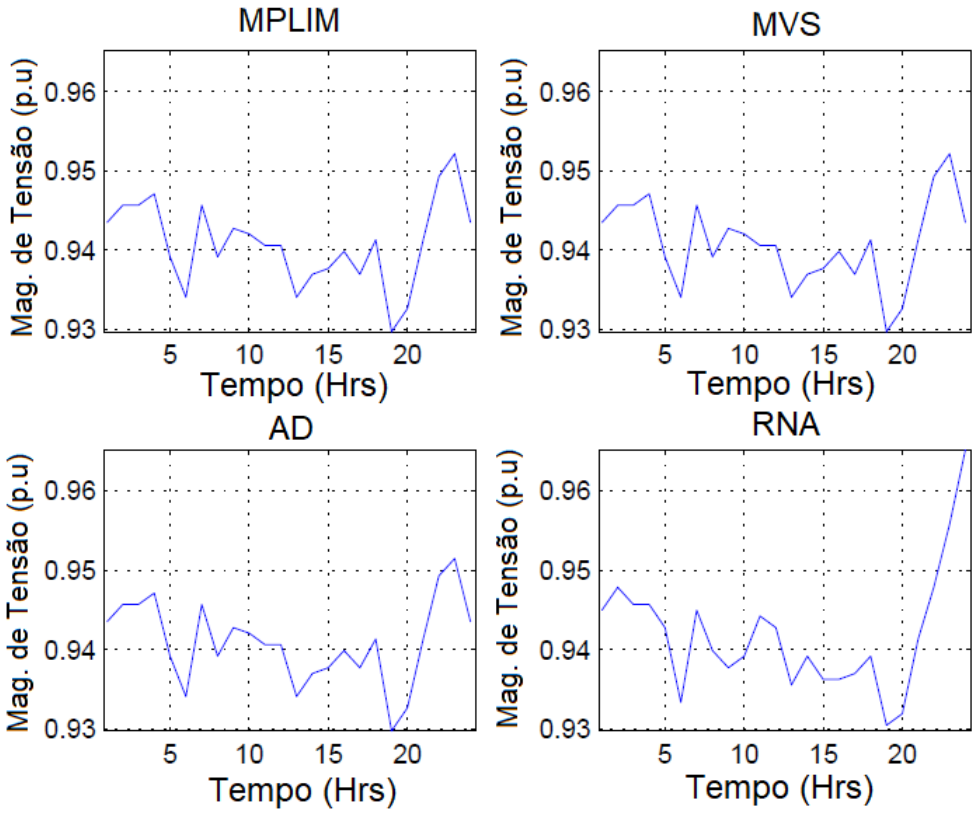
Figura 21 - Perdas de energia calculadas e ação de controle definida pelos SIs durante 24 horas.



Fonte: Próprio Autor

Na Figura 22, apresenta-se a magnitude de tensão mínima no sistema para cada intervalo de tempo (ação de controle definida pelos algoritmos que compõem os SIs durante 24 horas) para a percentagem de variação de demanda de 20%.

Figura 22 - Tensão mínima calculada e ação de controle definida pelos SIs durante 24 horas.



Fonte: Próprio Autor

Com cada base de dados usada para realizar o treinamento das ADs, das MVSs e das RNAs foram realizados testes nos quais foi necessário calibrar os diferentes parâmetros de cada um dos algoritmos. Um dos casos mais críticos aparece com as RNAs, as quais após vários testes no processo de treinamento, pode ser vista a dificuldade que existe na hora de encontrar o ajuste ótimo dos parâmetros de treinamento que são necessários quando se busca alcançar melhores resultados. Estes parâmetros são selecionados usando critérios de tempo e erro de treinamento. Para o caso das RNAs, os parâmetros mais relevantes são: camadas ocultas, iterações e classificação correta, portanto, foram realizados testes considerando variações na quantidade de camadas ocultas e quantidade de iterações visando obter uma melhor percentagem de classificação correta. Cabe salientar que o fato de ter maior quantidade de camadas ocultas ou de aumentar a quantidade de iterações não se vê diretamente relacionado com o aumento da percentagem de classificação correta, como pode ser observado na Tabela 3.

A calibração dos parâmetros dos SIs é de grande importância, pois, estão diretamente relacionados com a qualidade dos resultados obtidos.

Tabela 3 - Parâmetros utilizados para o treinamento da RNA

Camadas Ocultas	Iterações	Classificação Correta (%)
40	100	84,89
50	200	86,47
80	300	87,81
82	100	86,67
85	100	87,86
85	3000	94,91
90	100	86,55
100	400	87,12

Fonte: Dados da pesquisa do autor.

3.3.2 Validação dos SIs

No processo de validação dos SIs é dada total autonomia aos algoritmos , pois uma vez que já foram treinados, eles recebem os bancos de dados com as informações de entrada (medições), e posteriormente devolvem o controle definido pelo SI (saídas), seguidamente, este controle é comparado com o controle que realizou o MPLIM para verificar o nível de aprendizado dos SIs. Na Figura 23 é ilustrado o processo realizado na validação dos algoritmos que compõem os SIs. Com esta forma de validação, pretende-se avaliar a capacidade de aprendizagem e de generalização dos SIs ante o surgimento de novos cenários e de dados ruidosos.

Figura 23 - Processo de treinamento realizado entre o MPLIM e o SI.

Fonte: Próprio Autor

Para realizar a validação dos algoritmos que compõem os SIs, foram criadas quatro bancos de dados com o histórico de perdas de energia do sistema hora pós hora de 33 atributos durante uma semana fazendo uso do MPLIM. Cada base de dados contem 5544 cenários criados a partir de uma variação aleatória de demanda de 5%, 10%, 15% e 20% respetivamente.

Na Tabela 4, são comparados os resultados obtidos pelos SIs e os valores calculados pelo modelo MPLIM. Para ilustrar graficamente o processo de validação foi utilizada a base de dados que considerava a variação aleatória da demanda de 10%, com o objetivo de comparar o

desempenho de cada algoritmo que compõe o SI proposto para este caso em particular.

Tabela 4 - Perdas de energia em kWh obtidas durante uma semana usando as diferentes metodologias

	Dia 1	**Dia 2**	**Dia 3**	**Dia 4**	**Dia 5**	**Dia 6**	**Dia 7**
MPLIM	9767,97	9869,00	9916,46	9829,16	9868,82	9542,83	9618,72
MVS	9768,14	9891,00	9927,77	9855,30	9899,49	9633,30	9621,11
AD	9770,61	9878,02	9927,71	9867,10	9880,32	9564,92	9642,71
RNA	9767,97	9869,73	9958,95	9863,56	9883,45	9872,06	9763,50

Fonte: Dados da pesquisa do autor.

Assim, para validar os resultados escolhe-se desta base de dados, o dia 2 que é o dia que apresenta maiores níveis de perdas de energia. Na Tabela 5 é realizada uma comparação dos valores hora a hora das perdas de energia do sistema em estudo, obtidos pelo MPLIM e pelos três SIs para o dia que apresentam maiores níveis de perdas.

Tabela 5 - Valores calculados pelo MPLIM e definidos pelo SI das perdas do sistema hora a hora em kWh

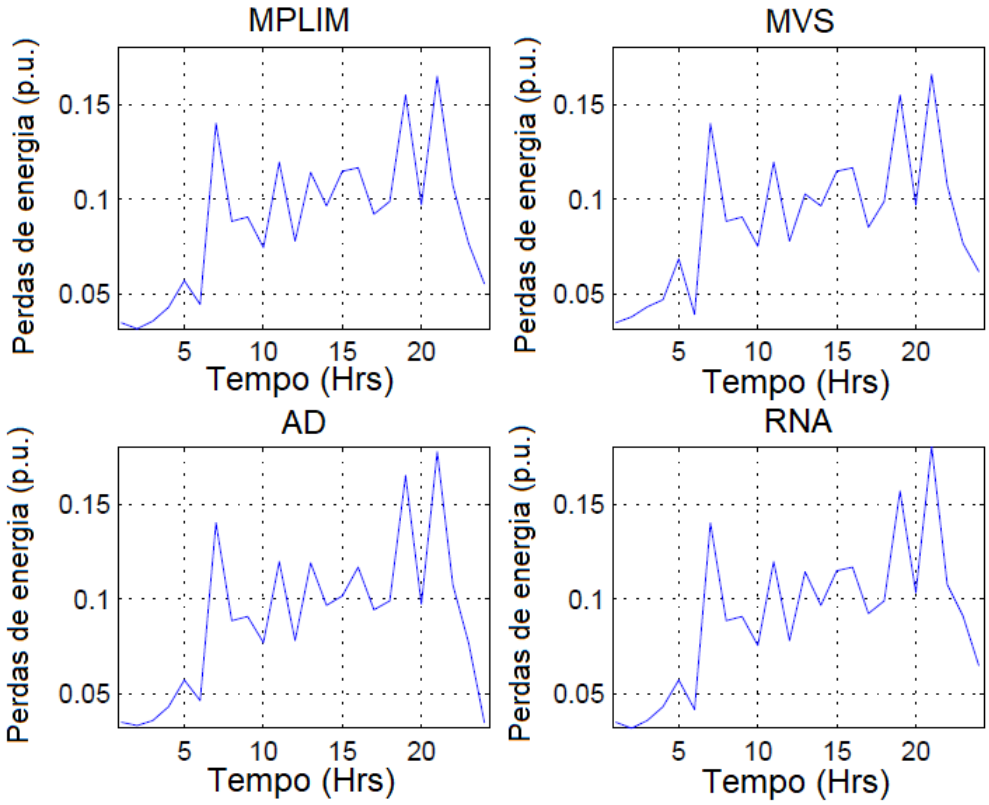
Perdas de energia em kWh				
	MPLIM	**MVS**	**AD**	**RNA**
Hora	**Dia 2**	**Dia 2**	**Dia 2**	**Dia 2**
0	152,93	152,93	152,93	152,93
1	137,62	137,62	137,62	137,62
2	134,00	134,00	134,00	134,00
3	122,58	122,58	122,58	122,92
4	177,18	177,18	177,18	177,18
5	229,29	228,12	228,35	228,35
6	424,63	424,63	424,63	424,63
7	543,70	566,87	543,91	543,70
8	487,85	487,85	487,85	487,85
9	511,15	511,15	511,15	511,15
10	505,39	505,39	505,39	505,39
11	476,51	476,51	476,51	476,51
12	652,43	652,43	652,43	652,43
13	600,66	600,66	600,66	600,66
14	580,35	580,35	581,99	580,35
15	540,70	540,70	540,70	540,70
16	547,93	547,93	547,96	547,93
17	478,44	478,44	478,44	479,77
18	662,89	662,89	662,89	662,89
19	596,88	596,88	596,88	596,88
20	464,58	464,58	464,58	464,58
21	360,00	360,00	360,00	360,00
22	321,58	321,58	329,66	321,58
23	159,73	159,73	159,73	159,73
	9869,00	**9891,00**	**9878,02**	**9869,73**

Fonte: Dados da pesquisa do autor.

Na Figura 24, apresentam-se as perdas de potência no sistema para cada intervalo de tempo (ação de controle definida pelo SIs durante 24 horas) para a percentagem de variação de demanda de 10%. Nesta figura, os valores das perdas obtidos usando a ação de controle ótima são os resultado obtidos na simulação do modelo MPLIM, e os outros, correspondem à ação

de controle definida pelos SIs. Nota-se a proximidade dos resultados previstos com relação aos assumidos como reais.

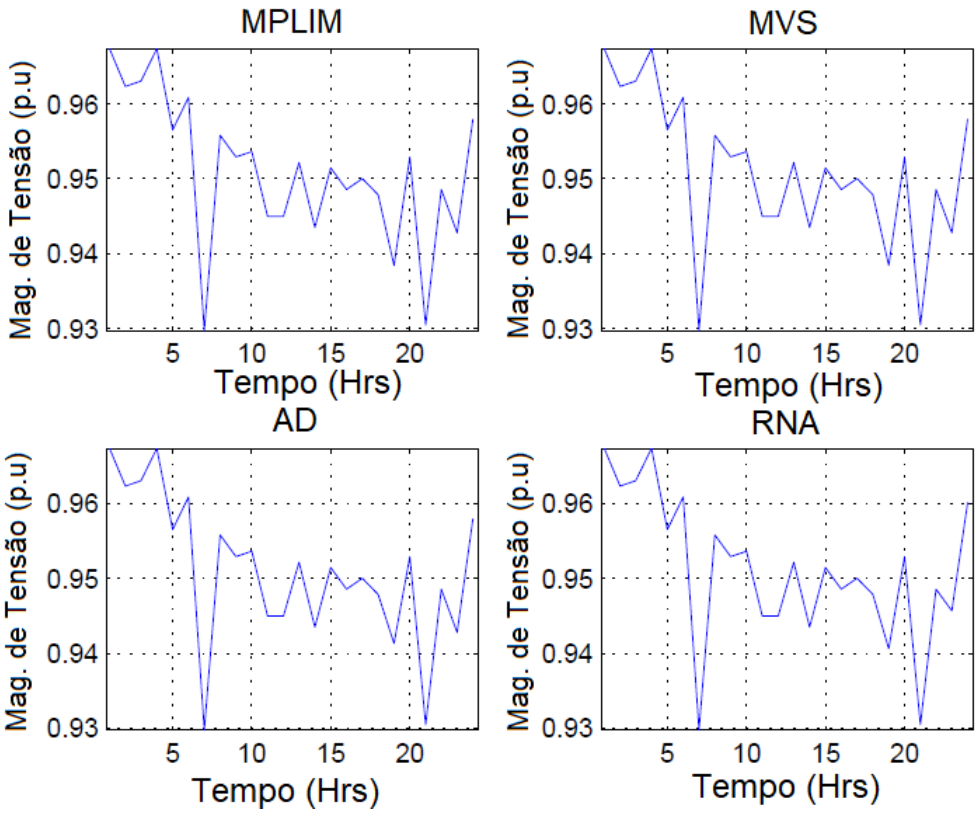
Figura 24 - Perdas de energia calculadas e ação de controle definida pelo SI durante 24 horas.



Fonte: Próprio Autor

Na Figura 25, apresentam-se a tensão mínima no sistema para cada intervalo de tempo (ação de controle definida pelo SIs durante 24 horas) para a percentagem de variação de demanda de 10%.

Figura 25 - Tensão mínima calculadas e ação de controle definida pelo SI durante 24 horas.



Fonte: Próprio Autor

Para melhor análise destas aproximações, são apresentadas tabelas com valores que permitem quantificar erros de previsão em termos de desvios médios absolutos (DMA).

Na Tabela 6, para o conjunto de validação com variação aleatória da demanda de 20%, apresentam-se os DMAs entre estados discretos, além das percentagens de acerto para cada um dos elementos de controle do sistema (GDs, BCs e RTs), quando são comparados os resultados de referência (resultados obtidos após a simulação do modelo MPLIM), com os resultados obtidos pelos SIs. De acordo com os resultados desta tabela, das metodologias dos SIs, pode-se notar que as MVSs apresentam um melhor desempenho, com maiores percentagens de desvios médios absolutos e maiores percentagens de acerto. Nesta tabela, também pode se observar que os menores desvios de acerto, tanto para as ADs, as MVSs, quanto para as RNAs, correspondem às previsões das potências injetadas pelos GDs; porém, com percentagens que podem ser considerados como aceitáveis. Novamente, pode-se afirmar que os resultados são aceitáveis, já que, para a maior percentagem de variação da demanda (20%), a percentagem de acertos continua sendo superior a 90% com os três SIs.

Nas Tabelas 7-11, apresentam-se o número de valores previstos usando o SI que tem apresentado melhores resultados, como é o caso das MVSs, estes valores são classificados de acordo com as percentagens de desvios médios absolutos quando comparados com os valores reais,

Tabela 6 - Percentagem de acertos e DMA por variável para 20% de variação na demanda

	MVSs		**AD**		**RNAs**	
	DMA entre Estados Discretos (%)	**Acertos (%)**	**DMA entre Estados Discretos (%)**	**Acertos (%)**	**DMA entre Estados Discretos (%)**	**Acertos (%)**
P1	0.6548	88.6905	0.9821	86.3095	1.2500	77.9762
P2	0.6845	87.5000	1.2798	83.3333	1.3095	79.1667
Q1	0.5357	89.2857	0.5357	89.2857	0.6845	86.9048
Q2	0.5060	89.8810	0.7143	85.7143	0.7143	85.7143
BC1	0.1190	97.6190	0.2381	95.2381	0.2083	95.8333
BC2	0.1786	97.0238	0.4464	92.8571	0.5060	91.0714
BC3	0.0000	100.0000	0.0000	100.0000	0.0000	100.0000
BC4	0.0893	98.2143	0.1786	96.4286	0.1786	96.4286
BC5	0.1190	97.6190	0.2679	94.6429	0.2976	94.0476
RT1	0.0000	100.0000	0.0000	100.0000	0.0000	100.0000
RT2	0.0000	100.0000	0.0000	100.0000	0.0000	100.0000
RT3	0.3571	92.8571	0.4167	91.6667	0.7738	85.7143
RT4	0.0000	100.0000	0.0000	100.0000	0.0000	100.0000
Total	**0.2495**	**95.2839**	**0.3892**	**93.4982**	**0.4556**	**91.7582**

Fonte: Dados da pesquisa do autor.

tanto para as tensões mínimas (em kV), quanto para as perdas do sistema.

Tabela 7 - DMA dos valores obtidos (5% de variação da demanda)

	0-20%	**20-40%**	**40-60%**	**60-80%**	**80-100%**
Tensão Mínima	161	2	1	0	4
Perdas	148	12	4	1	3

Fonte: Dados da pesquisa do autor.

Tabela 8 - DMA dos valores obtidos (10% de variação da demanda)

	0-20%	**20-40%**	**40-60%**	**60-80%**	**80-100%**
Tensão Mínima	139	4	4	0	11
Perdas	127	12	11	2	6

Fonte: Dados da pesquisa do autor.

Tabela 9 - DMA dos valores obtidos (15% de variação da demanda)

	0-20%	**20-40%**	**40-60%**	**60-80%**	**80-100%**
Tensão Mínima	127	1	6	0	13
Perdas	111	14	11	5	6

Fonte: Dados da pesquisa do autor.

Nota-se a alta quantidade de resultados previstos com desvios médios absolutos entre 0% e 20%, os quais são apresentados, também em percentagens, na Tabela 11. Desta forma, mostra-se a alta proximidade dos resultados previstos quando comparados com os valores de referência.

Tabela 10 - DMA dos valores obtidos (20% de variação da demanda)

	0-20%	20-40%	40-60%	60-80%	80-100%
Tensão Mínima	120	2	6	0	7
Perdas	103	10	10	4	8

Fonte: Dados da pesquisa do autor.

Tabela 11 - Percentagem de resultados previstos com valores de DMA entre 0% e 20%

	Variação 5%	Variação 10%	Variação 15%	Variação 20%
Tensão Mínima (%)	95.83	82.74	75.60	71.43
Perdas (%)	88.10	75.60	66.07	61.31

Fonte: Dados da pesquisa do autor.

Na Tabela 12 são apresentadas as percentagens de classificação corretas usando dados com variação aleatória da demanda (5% e 20%). Para estas percentagens os resultados obtidos mostram ser de boa qualidade.

Tabela 12 - Porcentagem de classificação usando Cross-Validation

Variação da demanda (5%)			Variação da demanda (20%)		
Classificação Correta (%)			**Classificação Correta (%)**		
AD	**MVS**	**RNA**	**AD**	**MVS**	**RNA**
94,60	94,87	91,25	94,15	94,67	91,36
94,70	95,08	92,30	94,31	94,81	91,26
94,67	95,15	92,04	94,28	94,86	91,49
96,86	95,06	92,17	94,11	94,87	91,00
94,74	98,45	93,49	94,16	95,00	91,43
94,09	94,22	91,49	94,20	94,85	91,25
94,10	94,23	91,40	94,34	94,76	91,55
94,20	94,24	90,26	94,42	94,87	91,26
96,16	94,25	91,41	94,34	94,88	91,26
94,10	94,31	91,24	94,27	95,02	91,37

Fonte: Dados da pesquisa do autor.

3.4 CONCLUSÕES DO CAPÍTULO

Foi usado o software livre WEKA para treinar e avaliar os três algoritmos de SIs (ADs, MVSs e RNAs) na definição dos pontos de operação de GDs, BCs e RTs.

A comparação entre os resultados obtidos na solução de um modelo MPLIM tendo como objetivo a minimização das perdas do sistema de distribuição, e os resultados obtidos usando os SIs levam a considerar o uso destas metodologias como alternativas autónomas, rápidas e eficientes.

Os resultados obtidos apresentam erros baixos em comparação com os resultados fornecidos por um modelo matemático, o que mostra a utilidade do método proposto.

De acordo com os resultados apresentados pelos métodos baseados em SIs, pode-se afirmar que, para as condições de teste consideradas, as MVSs apresentaram um melhor desempenho com menores percentagens de DMA e maiores percentagens de acerto. As menores percentagens de acerto, tanto para as ADs, as MVSs, quanto para as RNAs (para 20% de variação aleatória na demanda), correspondem às previsões das potências injetadas pelos GDs. Porém, com percentagens que podem ser considerados como aceitáveis.

Para os três SIs, realizaram-se testes para percentagens de variação aleatória da demanda (5%, 10%, 15% e 20%). Os desvios médios absolutos entre estados discretos e percentagens de acerto obtidos dos mencionados testes podem ser considerados aceitáveis já que, por exemplo, para a maior percentagem de variação da demanda (20%), a percentagem de acertos continua sendo superior a 90%.

O processo de tomada de decisão pode ser realizado sem precisar da representação detalhada do modelo físico da rede, pois, dados de medições podem ser utilizados como entrada para o treinamento do SIs;

Os tempos de treinamento dos SIs foram razoáveis, mas na hora de se realizar a avaliação dos SIs, os tempos de resposta no máximo de 1 segundo, tempo que é aceitável quando comparado com dispositivos utilizados para o controle de magnitude de tensão e de potência reativa (Volt-VAr) em tempo real.

4 DETECÇÃO DE FRAUDES NAS REDES DE DISTRIBUIÇÃO DE ENERGIA ELÉTRICA UTILIZANDO TÉCNICAS DE MINERAÇÃO DE DADOS

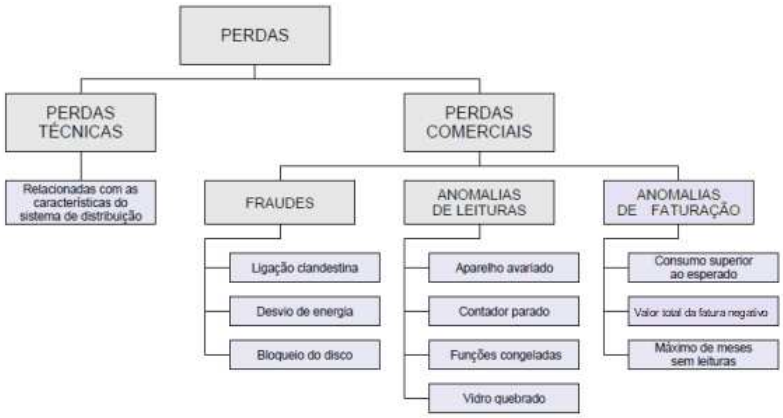
Neste capítulo é apresentada a aplicação de IC como solução ao problema de detecção de fraudes nos SDEEs utilizando MD. Este capítulo está organizado da seguinte forma: na Seção 4.1 é realizada uma introdução ao problema de detecção de fraudes. Posteriormente, são apresentados os diferentes métodos de solução encontrados na literatura para este problema. Logo após, na Seção 4.2 é apresentada a metodologia de solução proposta e na Seção 4.3 é apresentado o caso de estudo com seus resultados. Por último, na Seção 4.4 apresentam-se as conclusões.

4.1 INTRODUÇÃO

Um dos grandes problemas que frequentemente as concessionárias de energia elétrica enfrentam é a ocorrência de fraudes nos SDEEs.

No processo da distribuição da energia elétrica existem dois tipos de perdas de energia: as perdas técnicas e perdas comerciais (ANEEL, 2000). Estas perdas são uma questão fundamental do Setor de Distribuição. As perdas de energia elétrica, como já se infere pelo próprio nome, remetem à energia elétrica que, apesar de inserida no sistema interligado e na rede das distribuidoras, não chega a ser comercializada, seja por motivos técnicos ou por motivos de ordem comercial. Na Figura 26, são apresentadas as características dos diferentes tipos de perdas.

Figura 26 - Características das perdas.



Fonte: Próprio Autor

As **perdas técnicas** ocorrem naturalmente durante o processo de distribuição de energia

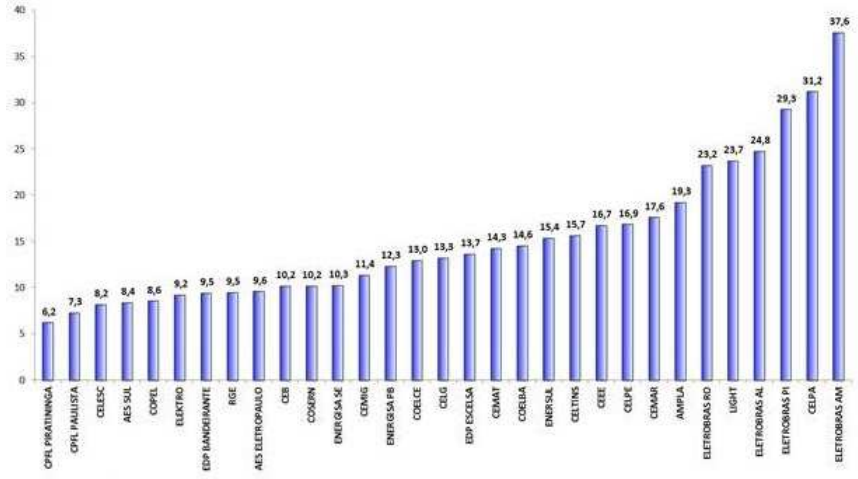
elétrica, isto é, estão relacionadas com as características próprias do sistema de distribuição. As perdas técnicas ocorrem por vários fenômenos elétricos, como é o caso do aquecimento dos fios condutores de energia, em decorrência da própria passagem da eletricidade, o chamado *Efeito Joule*. Nesse quesito, portanto, a extensão das redes e a grandeza territorial do Brasil acarretam impacto ao nível de perdas técnicas.

Em contra partida, as **perdas comerciais** estão relacionadas com as fraudes, "ligações clandestinas", falhas na medição, erros de leituras e faturamento, os quais fazem com que o desvio de energia elétrica da rede de distribuição passe diretamente para as instalações do consumidor sem passar pelo medidor de energia (ELLER, 2003). Este tipo de perdas, em geral, apresentam duas principais modalidades: furto e fraude de energia. O furto é caracterizado pelo desvio direto de energia da rede elétrica das distribuidoras para o consumidor ilegal.

Entende-se como furto o fato em que a energia é consumida e contabilizada pelas empresas concessionárias de energia elétrica (ECEE) mas não é realizado nenhum pagamento por parte do usuário, enquanto a empresa distribuidora continua tendo gastos derivados da distribuição desta energia que esta deixando de ser paga, o que leva às perdas comerciais. No caso das fraudes, as distribuidoras possuem um cadastro de todos os consumidores com as informações necessárias para realizar o controle do que deve ser faturado mensalmente, mas, usuários fraudulentos fazem adulterações no sistema de fiações elétricas da sua residência/comércio/indústria - de modo que, apesar de consumir uma quantidade X de energia, só paga efetivamente uma percentagem desse consumo, devido à fraude.

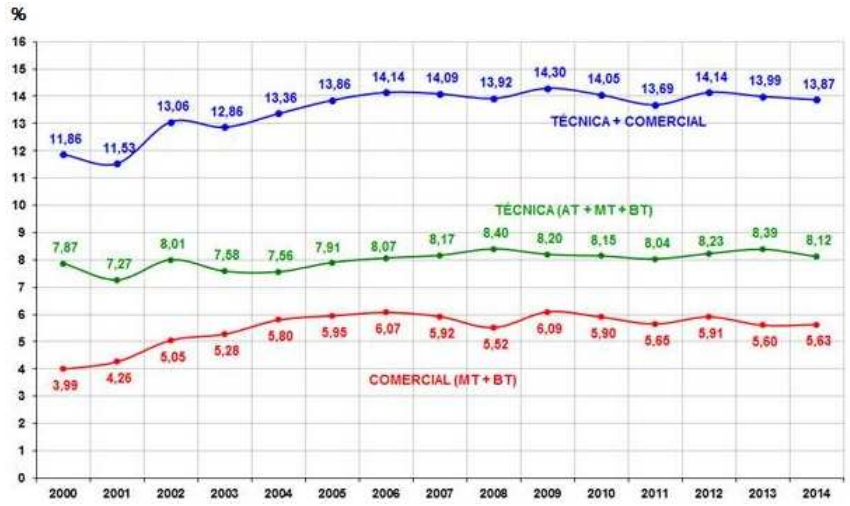
As Figuras 27 e 28 a seguir ilustram importantes estatísticas, elaboradas pela Associação Brasileira de Distribuidores de Energia Elétrica (ABRADEE) e parceiros, relacionadas às perdas de energia. De acordo com a ABRADEE, em 2014, as perdas totais em relação à energia injetada no sistema global das 64 distribuidoras foram de 13,87%, sendo que 8.12% correspondem às perdas técnicas e 5.63% correspondem às perdas não técnicas.

Figura 27 - Percentual de perdas de cada empresa distribuidora em 2014.



Fonte: Adaptado de SIG da ABRADEE (2016)

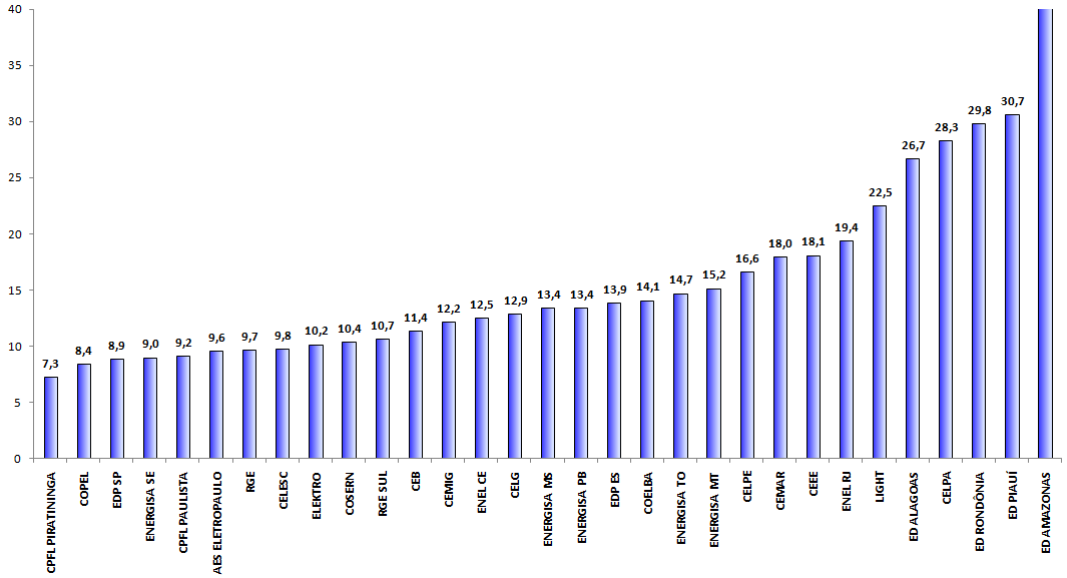
Figura 28 - Percentual de Perdas em Relação à Energia Injetada no Sistema Global das 64 Distribuidoras.



Fonte: Adaptado de SIG da ABRADEE (2016)

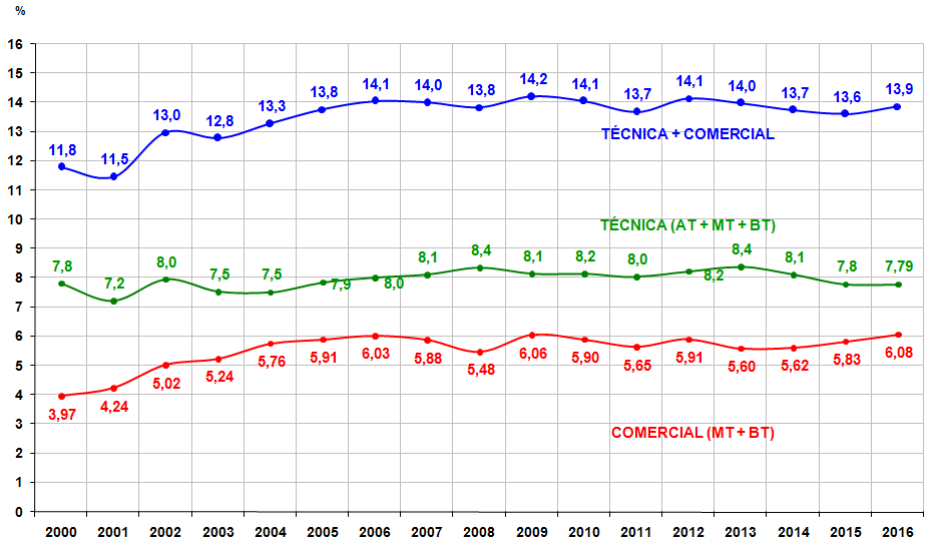
As Figuras 29 e 30 a seguir ilustram estatísticas, elaboradas pela ABRADEE e parceiros, relacionadas às perdas de energia. De acordo com a ABRADEE, em 2016, as perdas totais em relação à energia injetada no sistema global das 63 distribuidoras foram de 13.9%, sendo que 7.79% correspondem às perdas técnicas e 6.08% correspondem às perdas não técnicas.

Figura 29 - Percentual de perdas de cada empresa distribuidora em 2016.



Fonte: Adaptado de SIG da ABRADEE (2017)

Figura 30 - Percentual de Perdas em Relação à Energia Injetada no Sistema Global das 63 Distribuidoras.



Fonte: Adaptado de SIG da ABRADEE (2017)

Observando as Figuras 28 e 30, pode-se notar uma variação no percentual das perdas em relação à energia injetada no sistema global brasileiro até o ano 2016. Para o ano 2014 o percentual de perdas chegou a ser de 13,87%, em quanto, no ano 2016 chegou a ser de 13,9%. Poderia-se afirmar que não ouve muita variação segundo as percentagens apresentadas, mas na hora de analisar os diferentes tipos de perdas, surge a grande preocupação.

Até o ano 2014 as perdas técnicas representaram 8.12%, diferentemente do ano 2016 onde este tipo de perdas representou um percentual de 7.79%, este fato pode ser entendido como que ouve melhorias na parte técnica dos SDEE brasileiros. Em contra partida, tem-se que, até o ano 2014 as perdas comerciais representaram 5.63%, diferentemente do ano 2016 onde este tipo de perdas representou um percentual de 6.08%, portanto, ouve um aumento de 0.45%, isto é, a quantidade de fraudes tem aumentado.

Nas Figuras 27 e 29, pode ser visto que ouve um aumento notório na variação percentual de perdas entre os anos 2014 e 2016 em grande parte das ECEE do sistema global brasileiro. Segundo estas figuras, poucas empresas distribuidoras de energia elétrica têm conseguido diminuir o percentual de perdas.

Cabe lembrar que, mesmo entre as perdas não técnicas é necessário fazer uma distinção, pois existem as que não são gerenciáveis e as que podem ser gerenciadas. As perdas comerciais não gerenciáveis estão diretamente relacionadas com as ligações clandestinas feitas em lugares socialmente desassistidos, como favelas e palafitas. Neste sentido, é difícil determinar como pode haver uma distribuição de energia exemplar nestas localidades, em que as moradias estão totalmente ilegais. Perdas não técnicas relacionadas a condições desfavoráveis são casos que apresentam uma grande complexidade na hora de serem analisados.

As perdas gerenciáveis dizem respeito a fraudes nos medidores de energia elétrica e ligações clandestinas em regiões cujas condições de moradia não são tão desfavoráveis. Há também perdas não técnicas gerenciáveis ligadas a procedimentos errados e processos inadequados por parte da concessionária, uso de equipamentos obsoletos, falta de medidores e falta de rede. Outro problema é a utilização inadequada de medidores, isto é, quando um estabelecimento muda de perfil, passando de uma residência para um comércio, por exemplo, é necessário adequar o equipamento de medição para o nível de consumo adequado.

Dado que são várias as causas que podem gerar variação no montante que é faturado pelas ECEE ao que foi distribuído em certo período, tem-se apresentado inconsistências entre o valor medido e o valor calculado para o cálculo das perdas comerciais. Um dos fatores que altera a comparação entre estes dois valores é a impossibilidade de realizar a leitura de consumo de alguns usuários, o qual, provoca o faturamento realizado por estimativa, segundo o consumo dos clientes nos últimos meses (ABRADE, 2017).

Cabe salientar que, as principais causas de perdas comerciais segundo (ABRADE, 2017) são:

- Falta de medições.
- Falhas nos registros.
- Erros de medição e de faturamento.

- Fraude interna, isto é, erros de responsabilidade da empresa.
- Iluminação pública.
- Desvio de energia.
- Ligações clandestinas e fraudes.

Destas causas, as que representam maiores quantidades de perdas económicas para as ECEE são: o desvio de energia e as ligações clandestinas. Portanto, encontrar solução para estes problemas é de grande prioridade para ditas empresas, Delgado (2010).

Segundo (ARAÚJO, 2006), as estratégias mais usadas pelas ECEE para localizar ligações clandestinas e combater as fraudes são:

- Fiscalização de denúncias feitas pelos consumidores de forma anônima.
- Análises manuais das informações dos leitores de medidores e equipes de manutenção da rede.
- Atendimentos de emergências criadas por acidentes causados por certas ligações clandestinas.
- Inspeções nas unidades consumidoras.
- Identificação das áreas críticas.
- Balanço energético.
- Sistema de faturamento

Na atualidade, o mecanismo mais usado pelas empresas para identificar usuários fraudulentos, é mediante a realização de inspeções nas unidades consumidoras, mas, devido ao elevado número de unidades consumidoras que aparecem como suspeitas de fraudes, tais inspeções são realizadas sem uma análise eficiente do perfil de consumo dos clientes, o que representa um elevado custo para as empresas, (FILHO et al., 2004).

Cabe ressaltar que algumas empresas têm desenvolvido diferentes equipamentos que tornam a violação do medidor de energia elétrica uma tarefa mais difícil, o que é necessário em uma tentativa de reduzir as fraudes, Queiroga (2005). Além dos equipamentos sofisticados que estão sendo utilizados pelas distribuidoras para diminuir as perdas comerciais, pode-se notar que nos últimos anos, têm aumentado os investimentos na pesquisa e no desenvolvimento de métodos de detecção de fraudes através de sistemas computacionais inteligentes baseados na

técnica de IC, que permitem realizar análises dos dados dos consumidores, armazenados em grandes bases de dados (ELLER, 2003), (CABRAL et al., 2004), (FILHO et al., 2004).

Atualmente, os sucessivos incrementos nas contas de luz tem levado as ECEE a ficarem preocupadas com um possível crescimento da inadimplência e aumento nas ligações clandestinas. Esta preocupação está fundamentada no fato de que as perdas comerciais causadas por ligações irregulares não só representam perdas econômicas para empresa, como também criam um impacto social que tem que ser cuidadosamente analisado pelas ECEE.

Deve-se ter em conta que, as perdas comerciais são estudadas como um problema a nível nacional, dado que prejudica a sociedade, acarreta o aumento nas tarifas de fornecimento, sendo uma injustiça social. Mas estas, não são as únicas consequências que se tem pelas fraudes no SDEE, pois além do prejuízo financeiro das empresas, as ligaçoes irregulares podem causar acidentes fatais e incêndios. Isto ocorre pela alteração das características da rede. Portanto, as fraudes também representam um risco à segurança pública (DELGADO, 2010).

Segundo COPEL (2016), em 2016, as perdas globais, técnicas e não técnicas, representaram 9,6% da energia injetada no sistema da distribuidora, percentual que se manteve em relação ao ano 2015, como pode ser observado na Tabela 13. As perdas técnicas, nesta mesma base, permaneceram no patamar do ano anterior, e as perdas não técnicas apresentaram acréscimo de, aproximadamente, 0,4 p.p em 2016.

Tabela 13 - Comparação de perdas globais entre os anos 2014 e 2016.

Tipos de Perdas		2014	2015	2016
Transmissão	Perdas Globais (%)	1,7	1,8	2,0
	Perdas Técnicas (%)	1,7	1,7	1,9
Distribuição	Perdas Globais (%)	9,8	9,6	9,6
	Perdas Técnicas (%)	6,2	6,1	6,1
	Perdas Não Técnicas (%)	1,9	1,6	2,0

Fonte: Adaptado de COPEL (2016)

Segundo a superintendência de gestão tarifaria (ANEEL, 2015), a complexidade no combate às perdas não técnicas aumenta a cada dia, por se tratar de um problema que é, em alguma medida, impactado por aspetos socioeconômicos. Além disso, a detecção de possíveis fraudes bem como a sua prevenção configura-se um problema complexo. Mesmo que o histórico de consumo e o perfil do comportamento do consumidor apresentem claros indícios de fraude é importante que uma segunda investigação seja realizada. Portanto, as informações geradas pelos sistemas de apoio a decisão também precisam ser compatibilizadas com outras variáveis do sistema, para que cobranças erróneas não sejam aplicadas aos usuários.

4.2 METODOLOGIA PROPOSTA

A metodologia utilizada no desenvolvimento desta pesquisa é baseada em uma técnica de *IC*, chamada de DCBD mais comummente conhecida como MD. A decisão de utilizar a MD para resolver o problema de detecção de fraudes nos SDEEs, está baseada na sua capacidade de extração de conhecimento de grandes bases de dados, pois mediante o uso de algoritmos de AM, é possível realizar classificação dos usuários, detectar os tipos de irregularidades mais frequentes e potencialmente fraudulentas, e, assim, determinar o perfil do consumidor final, gerando estimativas sobre os consumidores suspeitos de cometer algum tipo de irregularidade.

Os algoritmos de AM utilizados no processo de MD, têm como objetivo principal interpretar os dados previamente selecionados a fim de produzir uma quantidade de padrões úteis, válidos e de fácil entendimento.

Os resultados gerados após o processo de MD, facilitam a tomada de decisões ágeis e inteligentes, a partir da extração de conhecimento e a análise estatística dos dados, permitindo que se observem padrões em situações em que se detecte o lugar onde as fraudes são mais frequentes e, classificação dos clientes quanto a confiabilidade, ou seja, determinar o perfil dos consumidores quanto aos diferentes tipos de anomalias apresentadas segundo um histórico de informações previamente analisadas.

Os algoritmos de AM escolhidos para resolver este problema foram: ADs, RNAs e MVS devido a suas características e suas capacidades de aprendizagem estudadas no Capítulo 2. Estes algoritmos são utilizados nos diversos processos de DCBD para possível detecção de fraudes nos SDEE. Uma vez escolhidos os algoritmos de AM que vão ser utilizados no processo de MD, o passo a seguir é escolher uma ferramenta eficiente para gerenciar de forma eficiente estes algoritmos. Segundo (LABERGE, 2011), para fazer está escolha é necessário ter em consideração o seguinte:

- Traduzir o problema da empresa a ser resolvido em séries de tarefas da MD.
- Compreender a natureza dos dados disponíveis em termos de conteúdo, tipos de atributos e estrutura das relações entre os atributos.

Para o gerenciamento dos algoritmos de AM utilizados nesta pesquisa, foi utilizado o software WEKA versão 3.8.0 para análise dos dados provenientes de um arquivo no formato ".csv", que foi gerado após o pré-processamento dos dados inicialmente entregues pela ECEE.

Como já foi estudado no Capítulo 2, a MD consta de três etapas: pré-processamento, MD e pós-processamento. Cada etapa foi considerada no desenvolvimento desta pesquisa e foi aplicada à solução do problema de detecção de fraudes da forma explicada a seguir.

4.2.1 Etapa de Pré-processamento

Durante o desenvolvimento desta etapa, foi necessário realizar algumas tarefas para o tratamento dos dados fornecidos pela ECEE. Inicialmente foi feita a limpeza dos dados com registros incompletos, ou com valores dos consumos fora dos limites estabelecidos. Outro tipo de tratamento dos dados foi a conversão do tipo dos dados, isto é, os atributos que compõem as bases de dados fornecidos que estão no formato de texto, continham dados de outros tipos, tais como, números inteiros, caracteres e datas.

Para realizar a construção da base de dados dos clientes a partir das bases de dados fornecidas pela ECEE, foram feitas as análises dos tipos de atributos que compõem a base de dados original a fim de selecionar um subconjunto de atributos, contendo informações realmente necessárias. O conjunto dos atributos que fazem parte da base de dados depois de realizar um processo de limpeza das bases de dados originais, está dividido em três grupos: informações do perfil dos usuários, histórico de consumo dos usuários e informações sobre as anomalias detectadas no processo de leituras do consumo de energia elétrica nos contadores. Salientando que estas são as principais informações utilizadas para realizar o pré-processamento dos dados.

Para atender as necessidades de análise das informações, foi necessário considerar as seguintes restrições:

- Um só tipo de usuário (residencial).
- Usuários com uma medição por mês .
- Usuários com todas as medições maiores que zero.
- Usuários com medições consideradas como válidas, isto é, sem erro de leitura.

Portanto, propõe-se o uso de um modelo de extração do conhecimento que avalia as características dos clientes (identificados pelo atributo MAT - Matricula) e, a partir delas, a criação de regras que serão aplicadas sobre os dados dos clientes finais para fazer uma redução do tamanho da base de dados, eliminando dados que não sejam relevantes no processo de MD. Uma das principais características a ser analisada é o histórico de consumo dos clientes (atributo CON_VAL - Consumo valorizado). Para isto, foram criadas quatro regras que avaliam os clientes com base nos seus históricos de consumo. Estas regras são:

- Classificação do consumo (CLA_CON).
- Risco (RIS).
- Oscilação (OSC).

- Anomalias (ANOM).

Essas regras são processadas de acordo com parâmetros de referência que indicam qual é a variação esperada para a classificação dos clientes. A seguir, apresenta-se a definição das regras anteriores.

- **Regra CLA_CON**

Os clientes são classificados de acordo com a potência consumida em: Consumo Baixo (CB), Consumo Médio (CM) e Consumo Alto (CA). A estrutura de classificação de consumo para CB, CM e CA é apresentada nas expressões CB, CM e CA, respectivamente.

- **CB** $\Rightarrow \frac{C_{\text{Max}}-C_{\text{Min}}}{3} \leq C(j) \leq \Rightarrow C_{\text{Min}}$

- **CM** $\Rightarrow \frac{C_{\text{Max}}-C_{\text{Min}}}{3} < C(j) \leq 2\frac{C_{\text{Max}}-C_{\text{Min}}}{3}$

- **CA** $\Rightarrow 2\frac{C_{\text{Max}}-C_{\text{Min}}}{3} < C(j) \leq C_{\text{Max}}$

Sendo $C(j)$ o consumo no mês j, e C_{Min} e C_{Max} os consumos mínimo e máximo especificados, respetivamente.

- **Regra RIS**

Usando um índice de risco, Ind_{Ris}, é determinada a possibilidade de fraude de acordo com a comparação da medição de consumo no mês atual, j, e a média das medições dos três meses anteriores. Nos três primeiros meses (j = 1, 2, 3) o índice de risco é igual a zero, caso contrário:

$$Ind_{\text{Ris}}(j) = \frac{C(j) - \frac{(C(j-3)+(C(j-2)+(C(j-1))}{3}}{\frac{(C(j-3)+(C(j-2)+(C(j-1))}{3}} * 100\% \tag{21}$$

Fazendo uso do índice de risco é determinado o nível de risco que tem cada cliente de estar cometendo fraude como Alto Risco (AR), Médio Risco (MR), Baixo Risco (BR) e Sem Risco (SR), de acordo com as expressões (22), (23), (24) e (25), respectivamente.

$$Ind_{\text{Ris}}(j) < -20\% \tag{22}$$

$$-20\% \leq Ind_{\text{Ris}}(j) \leq 0 \tag{23}$$

$$0 \leq Ind_{\text{Ris}}(j) < 20\% \tag{24}$$

$$20\% \leq Ind_{\text{Ris}}(j) \tag{25}$$

- **Regra OSC**

Usando um índice de oscilação, Ind_{Osc}, é determinada a possibilidade de fraude de acordo com a comparação da medição de consumo no mês atual, j, e a medição do mês anterior, $j-1$, sendo que para o primeiro mês (j = 1) o índice de mudança brusca é igual a zero, caso contrário:

$$Ind_{\mathrm{Osc}}(j) = \frac{C(j) - C(j-1)}{C(j-1)} * 100\% \tag{26}$$

Fazendo uso do índice de oscilação é classificado o cliente de acordo com o tipo de variação no consumo como de Oscilação Descendente (OsD), Oscilação Normal (OsN) e Oscilação Ascendente (OsA), de acordo com as expressões (27), (28), e (29), respectivamente.

$$Ind_{\mathrm{Osc}}(j) < -15\% \tag{27}$$

$$-15\% \leq Ind_{\mathrm{Osc}}(j) < 20\% \tag{28}$$

$$20\% \leq Ind_{\mathrm{Osc}}(j) \tag{29}$$

- **Regra ANOM**

Está regra é utilizada para determinar medições consideradas suspeitas e poderiam indicar usuários que devem ser inspecionados. A classificação correspondente a esta regra é determinada fazendo uso de um conjunto dos resultados obtidos da aplicação das regras CLA_CON, RIS e OSC. Com esta regra, podem ser definidos dois valores, como apresentado a seguir.

- Inspecionar (Ins): Se CLA_CON é CB, RIS é AR e OSC é OsD, ou, se CLA_CON é CB, RIS é MR e OSC é OsD, então é necessário inspecionar.
- Não Inspecionar (NIns): Em caso contrário aos mencionados anteriormente.

Os padrões de consumo de usuários com anomalias gerados na etapa de pré-processamento podem ser utilizados mais eficientemente, já que, tais dados, transformados em conhecimento, podem ser utilizados na tomada de decisões das empresas de forma sucinta.

4.2.2 Etapa de Mineração de Dados

A etapa de MD é realizada com o objetivo de extrair o conhecimento da base de dados e encontrar padrões de consumos com anomalias nas curvas de consumo dos usuários. Portanto,

deve ser considerado o fato de que os algoritmos de AM são treinados e posteriormente são utilizados para analisar novos dados.

Na hora de realizar o treinamento dos algoritmos de AM, são utilizados os dados considerados para extração do conhecimento, a fim de produzir padrões válidos e de fácil entendimento. Assim, são gerados resultados necessários para a tomada de decisões sobre possíveis anomalias ou fraudes. A etapa de MD é realizada fazendo uso de dados desde 2009 até 2013 para criar modelos de classificação e regressão, e uma percentagem destes dados foram armazenados para realizar a etapa de pós-processamento ou validação. Ao todo, após o pré-processamento e modelagem dos dados fornecidos pela empresa distribuidora, foram analisadas 112431 medições, correspondentes a 6474 clientes.

O treinamento dos algoritmos de AM é realizado com a base de dados modificada na fase de pré-processamento. Esta base de dados contem os resultados da análise das regras que avaliam o tipo de anomalias determinando se os usuários poderiam ou não ser fraudulentos.

A Figura 31 mostra imagens de algumas fraudes apresentadas por concessionárias de energia elétrica.

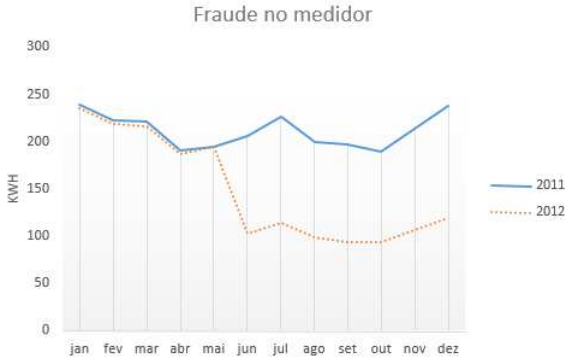
Figura 31 - Tipos de fraudes nas redes de DEE.

Fonte: Próprio Autor

A Figura 32 mostra graficamente outro tipo de padrão de fraude comum encontrado pelas concessionárias. É o caso dos fraudes nos medidores. Neste caso, a energia é medida em menor quantidade do que é inserida pela rede, ou seja, através de furos no disco do medidor, e/ou através da criação de dentes no disco, o medidor passa a errar na medida, sempre marcando

menos do que é realmente passado pelo medidor. Nota-se que, analisando o ano de 2011 e 2012, a curva de consumo mantém o mesmo padrão, apenas com uma redução de consumo a partir do mês de maio do ano de 2012.

Figura 32 - Alteração do medidor.



Fonte: Próprio Autor

Para identificar este tipo de anomalias e de possíveis fraudes foram utilizados algoritmos de AM. Um dos algoritmos de AM utilizados no processo de MD são as AD. O algoritmo de AD foi utilizado para identificador os tipos de irregularidades mais frequentes e potencialmente fraudulentas. As AD expressam uma forma simples de lógica condicional, selecionando subconjuntos baseados em valores para um determinado atributo.

As AD apresentadas nesta pesquisa, estão conformadas de nós internos que representam as perguntas de classificação e em suas folhas são apresentadas as partições do conjunto original ou seja a regra de classificação terá sempre no seu consequente uma resposta ao fato das condições satisfazerem ou não a uma determinada classe previamente definida.

A construção da AD objetiva predizer informações, gerando estimativas sobre os consumidores com risco de irregularidades e identificar padrões de comportamento.

No fase de treinamento, cada algoritmo de AM cria um modelo que é utilizado posteriormente no processo de avaliação na fase de pós-processamento.

4.2.3 Etapa de Pós-processamento

Esta é a fase final do processo de DCBD. Nesta fase são utilizados dados novos para avaliar os modelos de aprendizagem criados no processo de treinamento dos algoritmos utilizados para realizar a MD com o objetivo de medir o nível de aprendizado de cada algoritmo.

4.3 CASO DE ESTUDO E RESULTADOS

Como estudo de caso foi analisado o problema de fraudes em um SDEE colombiano administrado por uma ECEE que pretende diminuir o percentual de perdas comerciais existentes atualmente.

As bases de dados utilizadas nesta pesquisa contém dados reais que foram fornecidos pela ECEE. Inicialmente, uma das bases de dados continha um histórico de dados respetivos aos anos de 2009 e 2010, e a outra continha os dados dos anos de 2011 até 2013. No total, as bases de dados fornecidas por tal empresa consta de um histórico de 4.996.333 dados.

Estas bases de dados contém as seguintes informações: código do usuário, consumo, digito de checagem, o número da conta, o código do departamento (estado) e do município no qual o usuário mora, leitura dos contadores e dos transformadores, dois códigos relativos a atividades realizadas pelo cliente, classificação de serviço, nó e grupo de qualidade ao qual o usuário pertence, código do alimentador e do transformador usado pelo usuário, as fases utilizadas pelo usuário, faturação mensal e uma classificação do tipo de consumo (residencial de classe baixa, média ou alta). As duas bases de dados continham atributos que foram eliminados na fase de pré-processamento, já que não foram necessários no desenvolvimento desta pesquisa.

Como já foi mencionado, na Secção 4.2, pretende-se analisar esses dados com um software de livre distribuição, chamado WEKA. Com este software, busca-se realizar o processo de MD, dado que pesquisas realizadas em Witten e Frank (2000) demonstram que o WEKA tem a capacidade de fornecer resultados confiáveis e de qualidade.

4.3.1 Pré-Processamento dos Dados

A principal informação utilizada é o histórico de consumo de cada cliente. Além disso, o banco de dados está em um formato que não é aceito pelo WEKA, portanto, se deve transformar este banco de dados inicial para um banco de dados em formato específico tal como é o formato csv.

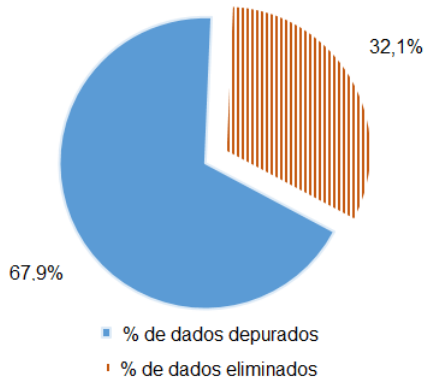
Para realizar as etapas de pré-processamento e transformação dos dados foi implementado um software utilizando a linguagem de programação JAVA. Esse software lê as bases de dados originalmente fornecidas pela ECEE, realiza uma junção das duas bases de dados, processa os dados aplicando as regras do modelo de extração e gera como resultado uma base de dados que contem os seguintes atributos:

- Um código único que é usado como identificador pelo sistema.
- Matrícula do usuário.

- Histórico de consumo entre os anos 2009 e 2013.
- O resultado da Classificação do Consumo (CLA_CON).
- O resultado do índice de Risco (RIS).
- O resultado da regra de Oscilação (OSC).
- O resultado da regra de Anomalia (ANOM).

Após a realização da etapa de pré-processamento dos dados, obtém-se uma redução dos dados contidos no banco de dados inicialmente estudado. Na Figura 33, é ilustrado o percentual de dados eliminados e dados depurados que fazem parte da nova base de dados criada durante a etapa de pré-processamento dos dados.

Figura 33 - Pré-processamento dos dados.



Fonte: Próprio Autor

Cabe salientar que, esta nova base de dados contém as informações necessárias para a realização da etapa de MD.

4.3.2 Mineração de Dados

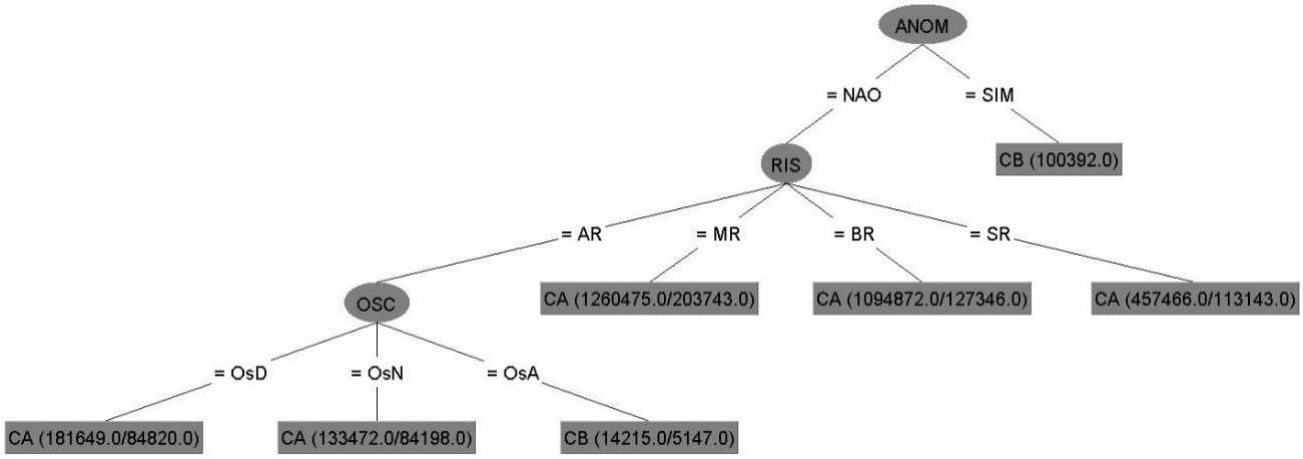
Nesta etapa, propõe-se o uso de modelos de extração do conhecimento que avalia as características dos clientes para criar um padrão a partir da informação contida no banco de dados. Para tal, realizam-se análises dos dados através dos algoritmos de AM considerados para a MD a fim de produzir padrões válidos e de fácil entendimento. Cada um dos algoritmos foi testado separadamente na fase de MD e os resultados são apresentados a seguir.

4.3.2.1 Aplicação do Algoritmo de AD J48

O algoritmo de AD J48 foi utilizado para realizar a fase de MD. Com este algoritmo é possível realizar a análise dos dados a partir da divisão dos mesmos fazendo uso de um atributo contido na base de dados.

Para o primeiro teste realizado foi considerado o atributo *CLA_CON* e os resultados obtidos com o uso deste algoritmo foram a matriz de confusão da Tabela 14 e a AD da Figura 34. Da matriz de confusão obtida, observa-se que considerando o atributo CLA_CON 2.624.144 instâncias de um total de 3.242.541 foram classificadas corretamente (80,92% do total de dados) para os valores CM, CB e CA, respectivamente, enquanto que 477.447, 140.191 e 759 instâncias (19,08% do total dos dados) foram classificadas incorretamente para os mesmos valores. Em uma matriz de confusão, quanto maiores os valores da diagonal principal, melhor a classificação obtida. Neste caso, note-se que somente para os valores OsN e OsA isto é satisfeito. De acordo com as percentagens de classificação obtidas para o atributo CLA_CON, pode-se deduzir que há confiabilidade "média" nesta classificação.

Figura 34 - Árvore de Decisão (Atributo CLA_CON).



Fonte: Próprio Autor

Tabela 14 - Matriz de Confusão (Atributo OSC).

	CM	CB	CA
CM	0	4388	473059
CB	0	109460	140191
CA	0	759	2514684

Fonte: Dados da pesquisa do autor

A informação proporcionada na AD da Figura 34 permite realizar as seguintes observações:

- Se CLA_CON tem o valor de CA e, OSC tem o valor de OsD e, RIS tem o valor de AR,

então a classificação propõe que em 84.820 de 181.649 medições não existem anomalias. Portanto para esses valores não são sugeridas inspeções.

- Se CLA_CON tem o valor de CA e, OSC tem o valor de OsN e, RIS tem o valor de AR, então a classificação propõe que em 84.198 de 133.472 medições não existem anomalias. Portanto para esses valores não são sugeridas inspeções.

- Se CLA_CON tem o valor de CB e, OSC tem o valor de OsA e, RIS tem o valor de AR, então a classificação propõe que em 5.417 de 14.215 medições não existem anomalias. Portanto para esses valores não são sugeridas inspeções.

- Se CLA_CON tem o valor de CA e, RIS tem o valor de MR, então a classificação propõe que em 203.743 de 1.260.475 medições não existem anomalias. Portanto para esses valores não são sugeridas inspeções.

- Se CLA_CON tem o valor de CA e, RIS tem o valor de BR, então a classificação propõe que em 113.143 de 127.346 medições não existem anomalias. Portanto para esses valores não são sugeridas inspeções.

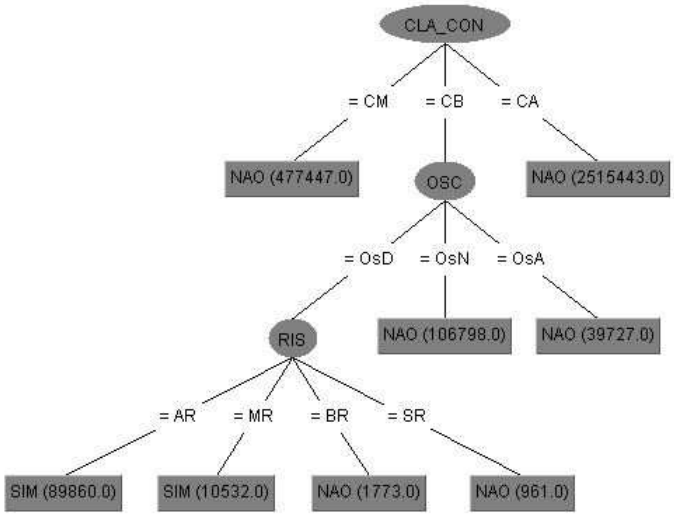
- Se CLA_CON tem o valor de CA e, RIS tem o valor de SR, então a classificação propõe que em 457.466 de 1.097.872 medições não existem anomalias. Portanto para esses valores não são sugeridas inspeções.

- Se CLA_CON tem o valor de CB, então a classificação indica que 100392 medições existem anomalias. Portanto para esses valores são sugeridas inspeções.

Para o segundo teste realizado foi considerado o atributo *ANOM* e os resultados obtidos com o uso deste algoritmo são mostrados adiante.

A matriz de confusão da Tabela 15 e a AD da Figura 35 foram obtidas aplicando o algoritmo J48 considerando o atributo ANOM. Nos testes realizados, esta classificação é a que proporcionou maior percentagem de acerto, 100%. Da matriz de confusão obtida, observa-se que para este atributo 3.242.541 instâncias do total de 3.242.541 foram classificadas corretamente. De acordo com as percentagens de classificação obtidas considerando o atributo ANOM, pode-se deduzir que há confiabilidade “alta” nesta classificação.

Figura 35 - Árvore de Decisão (Atriibuto ANOM).



Fonte: Próprio Autor

Tabela 15 - Matriz de Confusão (Atributo ANOM).

	Não	Sim
Não	3142149	0
Sim	0	100392

Fonte: Dados da pesquisa do autor

De acordo com a AD da Figura 35, as seguintes observações podem ser feitas:

- Se CLA_CON tem o valor CB e, OSC tem o valor OsD e, RIS tem o valor AR, então a classificação indica que 89.860 medições são suspeitas de possuir algum tipo de fraude, portanto sugerem inspeção.
- Se CLA_CON tem o valor CB e, OSC tem o valor OsD e, RIS tem o valor MR, então a classificação indica que 10.532 medições são suspeitas de possuir algum tipo de fraude, portanto sugerem inspeção.
- Se CLA_CON tem o valor CB e, OSC tem o valor OsD e, RIS tem o valor BR, então a classificação indica que 1.773 medições não são suspeitas de fraudes.
- Se CLA_CON tem o valor CB e, OSC tem o valor OsD e, RIS tem o valor SR, então a classificação indica que 961 medições não são suspeitas de fraudes.
- Se CLA_CON tem o valor CB e, OSC tem o valor OsN, então a classificação indica que 106.798 medições não são suspeitas de fraudes.

- Se CLA_CON tem o valor CB e, OSC tem o valor OsA, então a classificação indica que 39.727 medições não são suspeitas de fraudes.
- Se CLA_CON tem o valor CM, então a classificação indica que 447.447 medições não são suspeitas de fraudes.
- Se CLA_CON tem o valor CA, então a classificação indica que 2.515.443 medições não são suspeitas de fraudes.

Além destes atributos foram testados os atributos de *Oscilação* e *Risco* utilizando os algoritmos de AD, mas o atributo que obteve o melhor resultado foi o atributo *Anomalia*, por esse motivo foi dada maior relevância a esse teste nesta pesquisa.

4.3.2.2 Aplicação do algoritmo de MVS

Uma vez treinados os algoritmos de AD na fase de MD, é realizado o processo de MD aplicando os algoritmos de MVS.

Os treinamentos realizados empregando o algoritmos de MVS permitiram modificações em diversos parâmetros. No algoritmo LibSVM, alguns dos atributos calibrados foram: custo, degrau, Kernel, entre outros. E no algoritmo SMO foram: Kernel, Complexidade C, Filtro dos dados, entre outros. Porém as alterações dos parâmetros não mudaram a taxa de acerto, levando a se obter mudança apenas nos tempos gastos durante o treinamento e validação.

A porcentagem de acerto obtida utilizando os algoritmos de MVS foi de 100%. Tanto no algoritmo LibSVM quanto no SMO não foram encontrados padrões diferente ao mudar as configurações do algoritmo. As Tabelas 16 e 17 exibem os resultados obtidos no processo de teste dos algoritmos LibSVM e SMO respectivamente.

Comparando as informações contidas nas Tabelas 16 e 17, relacionadas aos algoritmos de MVS, pode-se observar que o atributo que tem maior influência sob o resultado é o Kernel. Este atributo é o responsável pela formulação matemática do problema a ser solucionado. Em relação ao algoritmo LibSVM, observa-se que o Kernel de maior velocidade -em quanto ao processamento de dados- é o Linear, e o mais lento é o Polynomial. Já no algoritmo SMO, o Kernel de maior velocidade é o Polynomial, enquanto ao mais lento é o Radial Basis Function (RBF).

Também pode-se afirmar que em média, os resultados do algoritmo LibSVM possuem o treinamento e testes mais rápidos quando comparado aos resultados do SMO, porém, se for observado o melhor teste de cada um dos algoritmos aqui tratados, o algoritmo SMO é quem possui o melhor tempo de treinamento do modelo. Em outras palavras, pode-se afirmar que o Kernel Polynomial aplicado em conjunto com o algoritmo SMO possui curto tempo de treino,

Tabela 16 - Comparação dos resultados obtidos do algoritmo LibSVM.

Kernel	Degre	Coef	Custo	Instancias de treinamento	Instancias de Teste	Tempo de treinamento (seg)	Tempo de validação (seg)	Acertos (%)
RBF	3	0	1	2.918.287	324.254	71,4	3,87	100
Linear	3	0	1	2.918.287	324.254	25,93	2,65	100
Polynomial	3	0	1	2.918.287	324.254	71,54	6,58	100
Sigmoid	3	0	1	2.918.287	324.254	77,61	4,65	100
Linear	3	0	5	2.918.287	324.254	24,64	2,47	100
Linear	3	0	15	2.918.287	324.254	27,59	2,36	100
Polynomial	3	0	2	2.918.287	324.254	52,25	4,2	100
Polynomial	3	0	3	2.918.287	324.254	45,93	3,43	100
Polynomial	3	0	4	2.918.287	324.254	42,34	3,28	100
Polynomial	3	2	3	2.918.287	324.254	62,17	2,48	100
Polynomial	3	4	5	2.918.287	324.254	96,83	2,37	100
Polynomial	1	0	1	2.918.287	324.254	41,33	3,26	100
Polynomial	3	0	1	2.918.287	324.254	114,6	8,52	100
Polynomial	5	0	1	2.918.287	324.254	535,27	44,304	100
Polynomial	1	0	1	2.918.287	324.254	40,06	3,151	100
Polynomial	5	0	1	2.918.287	324.254	535,27	44,304	100
Sigmoid	3	0	10	2.918.287	324.255	27,02	1,966	100
Sigmoid	3	0	30	2.918.287	324.254	55,08	2,824	100
RBF	3	10	1	2.918.287	324254	69,19	2,517	100
RBF	3	30	1	2.918.287	324254	66,67	2,665	100
RBF	10	0	1	2.918.287	324254	66,6	3,11	100
RBF	20	0	1	2.918.287	324254	68,05	2,775	100
RBF	15	8	10	2.918.287	324254	66,27	2,605	100
Linear	3	0	25	2.918.287	324.254	25,33	2,469	100
Linear	3	0	50	2.918.287	324.254	30,8	2,201	100

Fonte: Dados da pesquisa do autor

Tabela 17 - Comparação dos resultados obtidos do algoritmo SMO.

Complexidade C	Kernel	Filtro	Instancias de Treino	Instancias de Teste	Tempo de treino (seg)	Tempo de teste (seg)	Acertos
0,1	Puk	Normalize	2.918.286	324.255	467,52	14,82	100
0,2	Polynomial	Normalize	2.918.286	324.255	40,91	11,16	100
0,5	Normalized Poly Kernel	Normalize	2.918.286	324.255	225,56	14,28	100
0,8	Normalized poly Kernel	Normalize	2.918.286	324.255	302,34	16,99	100
2	Normalized poly Kernel	Normalize	2.918.286	324.255	167,41	13,84	100
0,7	Polynomial	Normalize	2.918.286	324.255	20,48	10,11	100
1	Poly Kernel	Normalize	2.918.286	324.255	20,58	10,48	100
1	Normalized poly Kernel	Normalize	2.918.286	324.255	280,25	6,62	100
1	Puk	Normalize	2.918.286	324.255	386,63	18,02	100
1	RBF Kernel	Normalize	2.918.286	324.255	1201,84	32,95	100
1	RBF Kernel	Standardize	2.918.286	324.255	1229,47	31,81	100
1,5	Puk	Normalize	2.918.286	324.255	431,7	18,77	100
2	Normalized Poly Kernel	Normalize	2.918.286	324.255	171,98	14,71	100
3	RBF Kernel	Normalize	2.918.286	324.255	801,31	22,46	100
6	RBF Kernel	Normalize	2.918.286	324.255	634,53	23,28	100
6	RBF Kernel	Standardize	2.918.286	324.255	870,45	22,51	100
8	Puk	Normalize	2.918.286	324.255	420,31	18,58	100
8	RBF Kernel	Standardize	2.918.286	324.255	1265,44	33,32	100
10	RBF Kernel	Normalize	2.918.286	324.255	1010,34	23,88	100
20	Normalized Poly Kernel	Normalize	2.918.286	324.255	167,41	13,84	100
20	Normalized poly Kernel	Standardize	2.918.286	382.255	189,37	15,53	100
20	RBF Kernel	Normalize	2.918.286	324.255	1129,67	33,57	100
50	Polynomial	Normalize	2.918.286	324.255	20,75	10,91	100

Fonte: Dados da pesquisa do autor

mas é lento ao avaliar o modelo, em contrapartida, o Kernel Linear quando aplicado em conjunto com o algoritmo LibSVM consome bastante tempo no treinamento, porém é mais veloz que todos os outros Kernels no momento de avaliar o modelo criado.

Um outro teste realizado para verificar a capacidade de aprendizagem das MVS, foi realizado fazendo uma divisão da base de dados, para utilizar o histórico de dados de forma anual. O objetivo deste teste era observar o comportamento do processo de MD quando se tem um histórico de dados consideravelmente pequeno.

Neste teste foi realizado o processo de MD utilizando a base de dados dos anos 2009 e 2010. Uma vez criado o modelo de treinamento é realizada a validação do mesmo fazendo uso de novas bases de dados. Para este caso foi utilizada uma base de dados do ano 2011 e de 2012 para realizar a validação em dois momentos diferentes com cada uma das bases de dados.

Os resultados obtidos no processo de treinamento e validação são os seguintes: no processo de treinamento o modelo criado a partir do algoritmo de SMO de MVS alcançou um acerto de 96,67%, enquanto que, para os dados correspondentes aos anos de 2011 e 2012 se obteve um resultado com uma pequena variação entre a percentagem de acerto. A Tabela 18 mostra a percentagem de acerto durante o treinamento com os dados de 2009 e 2010, e os testes de validação com os dados de 2011 e 2012.

Tabela 18 - Testes com o algoritmo de MVS

	acerto	erro
Treinamento do Modelo	96,67%	3,33%
Validação - dados de 2011	96,74%	3,26%
Validação - dados de 2012	98,24	1,76%

Fonte: Dados da pesquisa do autor.

Segundo este teste, pode-se afirmar que o tamanho da base de dados que contem o histórico de dados de um usuário, pode sim afetar o resultado do processo de MD.

4.3.2.3 Aplicação do algoritmo de RNA

O terceiro algoritmo proposto para resolver este problema de detecção de fraudes nas redes de distribuição de energia elétrica através da aplicação da MD são as RNAs. A seguir é realizado o processo de MD aplicando o algoritmo de RNA, com o objetivo de comparar as diferenças entre a capacidade e eficiência dos diferentes algoritmos de AM utilizados no processo de MD nesta pesquisa.

Aplicando o algoritmo de RNA e alterando as configurações do algoritmo em busca de um melhor resultado no treinamento, foi possível observar que a alteração nas configurações que

mudou a porcentagem de acerto foi somente o *TrainingTime*. Essa configuração representa o número de iterações necessárias para treinar completamente a rede, e a porcentagem de acerto de maior valor alcançou 92,7324%, como pode ser observado na Tabela 19.

Tabela 19 - Parâmetros do algoritmo de RNA

Teste	*TrainingTime (iterações)*	Camadas Ocultas	Acerto
01	2.000	3	80,5422%
02	5.000	10	84,3011%
03	10.000	15	86,4733%
04	50.000	30	91,0249%
05	100.000	5	92,7324%
06	200.000	5	92,6829%

Fonte: Dados da pesquisa do autor

É importante salientar que na Tabela 19 só foram apresentados os dados relacionados com os atributos do algoritmo de RNA que, após as alterações na configuração, obtiveram alterações notáveis nos resultados durante o treinamento da rede.

Vale ressaltar que as modificações de aumento no número de iterações, quando chega a um determinado valor, começa a ter queda no resultado, como pode ser observado no teste número 06 apresentado na Tabela 19.

O modelo RNA que alcançou 92,7324% de acerto durante o treinamento do modelo, classificou os 314.023 usuários de 2011 e os 313.972 usuários de 2012 durante o processo de validação do modelo, com um tempo de execução para classificação de 4 segundos, e a porcentagem de acerto na classificação foi de 92,7324% em ambas as classificações, ou seja, da base de dados inserida com as informações dos usuários, 92,7324 % foram classificados corretamente, sendo que esta classificação permite observar que usuários podem ou não ser suspeitos de realizar algum tipo de anomalias ou fraudes de consumo de energia elétrica. A Tabela 20 mostra a porcentagem de acerto na classificação durante o modelo e os dois anos classificados.

Tabela 20 - Testes com o modelo RNA

	acerto	erro
Treinamento do Modelo	92,73%	7,27%
Validação - dados de 2011	92,73%	7,27%
Validação - dados de 2012	92,73%	7,27%

Fonte: Dados da pesquisa do autor

É importante ressaltar que o processo de treinamento é lento. Alguns treinamentos demoram cerca de 15 dias para se obter um modelo de boa qualidade. Mas o tempo de treinamento,

como é realizado apenas para a criação do modelo, não prejudica na utilização de MD para classificações, pois depois do modelo pronto, a execução para classificar novos dados, variaram de 3 a 10 segundos com os dados e modelos utilizados no presente trabalho.

4.3.3 Pós-processamento

No processo de pós-processamento é realizada a comparação dos resultados obtidos com resultados entregues pela ECEE. Este processo foi possível graças a que a empresa que forneceu os dados para realizar a pesquisa já tinha feito a detecção de usuários suspeitos de fraudes, e esta informação permitiu validar se os resultados obtidos na pesquisa realmente estavam em concordância com os resultados obtidos de forma manual pela empresa.

Os resultados obtidos durante o desenvolvimento desta pesquisa, além de ter detectado os mesmos usuários fraudulentos que a ECEE já tinha detectado, também foram encontrados outros usuários suspeitos de fraude que a ECEE não tinha inspecionado. Com as novas informações a ECEE conseguiu realizar inspeções nas unidades consumidoras de forma mais eficiente, dado que normalmente este processo de detecção de usuários suspeitos de fraude é realizado de forma manual por um ou mais trabalhadores da ECEE em questão.

4.4 CONCLUSÕES DO CAPÍTULO

A comparação entre os resultados obtidos na detecção de fraudes de forma manual tendo como objetivo a minimização das perdas não técnicas, e os resultados obtidos usando MD levam a considerar o uso destas metodologias como alternativas autônomas, rápidas e eficientes.

Observou-se que aplicando o algoritmo de ADs no processo de MD considerando a relevância do atributo ANOM, foi obtida a maior percentagem de acerto na classificação (100%), representando uma alta confiabilidade na tomada de decisões a partir do uso desta metodologia.

De acordo com os resultados apresentados pelo método de MD, pode-se afirmar que, para as condições de teste consideradas, as ADs apresentaram um melhor desempenho com percentagens de acerto de (100%). As menores percentagens de acerto, tanto para as MVSs, quanto para as RNAs, correspondem a percentagens que podem ser considerados como aceitáveis.

Verificou-se que através da MD é possível realizar análises para a deteção de possíveis fraudes nas ECEE. Além disso, demonstrou-se que a etapa de pré-processamento dos dados é de grande importância para a obtenção de resultados confiáveis.

Com os resultados obtidos após a aplicação dos algoritmo de MD foi possível cumprir com o objetivo desta pesquisa que é classificar se determinado consumidor de energia pode ou não ser suspeito de estar realizando algum tipo de fraude, além disso, foi realizada a tarefa mais

adequada dentro do processo de MD para o processo de classificação.

Foi usado o software livre WEKA para treinar e avaliar os três algoritmos de AM (ADs, MVSs e RNAs) utilizados no processo de MD, na detecção de fraudes nos SDEE.

Os tempos de treinamento dos algoritmos de AM utilizados na etapa de MD foram razoáveis, mas na hora de se realizar a avaliação dos SIs, os tempos de resposta foram no máximo de 383 segundos, tempo que pode ser considerado aceitável quando comparado com o processo de detecção de fraudes de forma manual.

5 LOCALIZAÇÃO DE FALTAS EM LINHAS DE TRANSMISSÃO DE ENERGIA ELÉTRICA APLICANDO ALGORITMOS DE INTELIGÊNCIA COMPUTACIONAL

Neste capítulo é apresentada a aplicação de IC como solução ao problema de localização de faltas em STEE. Este capítulo está organizado da seguinte forma: na Seção 5.1 é realizada uma introdução ao problema e, posteriormente, na Seção 5.2 são apresentados os diferentes métodos de solução encontrados na literatura. Seguidamente, na Seção 5.3 é apresentada a metodologia de solução proposta e na Seção 5.4 é apresentado o caso de estudo com seus resultados. Por último, na Seção 5.5 apresentam-se as conclusões do capítulo.

5.1 INTRODUÇÃO

As redes inteligentes têm atraído recentemente a atenção de muitos grupos de pesquisa com a sua capacidade para criar uma automática e distribuída entrega de energia. A IC foi incorporada em vários aspectos às redes inteligentes, incluindo detecção de faltas e classificação, que é uma questão-chave em todos os sistemas de energia.

Redes inteligentes como um sistema, usam IC de forma integrada em toda a geração, transmissão, distribuição e consumo para alcançar a segurança, confiabilidade, resiliência e eficiência. Em contraste com a rede elétrica tradicional, uma rede inteligente utiliza fluxos bidirecionais de informações e eletricidade para criar uma rede de distribuição de energia avançada, automatizada e distribuída (FANG et al., 2012). Sistemas de geração distribuída, de duas vias de comunicação segura, auto monitorização e auto recuperação são algumas características que diferenciam as redes inteligentes das redes tradicionais de energia (GHARAVI; GHAFURIAN, 2011).

Certamente a ideia de aplicar a inteligência de máquina e aprendizagem de máquina aos sistemas de energia elétrica não é nova. De fato, quase todas as áreas relacionadas com as redes de energia têm recebido atenção da pesquisa em IC (Saxena, 2010). Isso inclui operação de sistemas elétricos de potência, planejamento, controle, mercados, automação, aplicações em sistemas de distribuição incluindo GD, e previsão do tempo. São muitas as pesquisas desenvolvidas na área de IC aplicada nas redes de energia elétrica. No entanto, todas estas aplicações de IC têm sido desenvolvidas para partes específicas da rede em funcionamento. Nesta tese foi pesquisada a aplicação de algoritmos de IC na resolução de problemas tais como a localização de faltas em STEE.

Um dos aspectos mais importantes das redes inteligentes é a sua aplicação na detecção e

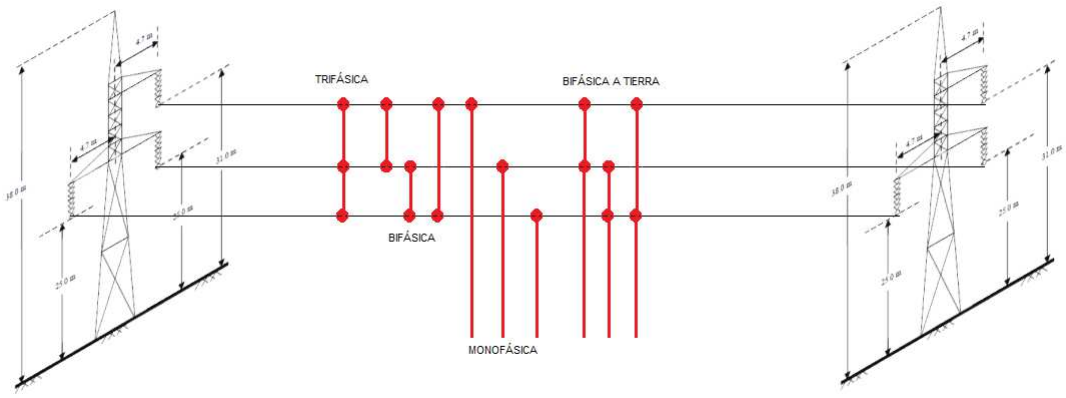
classificação de faltas, e análise de rotina de perturbações subjacentes que levam a faltas. Contudo, para que uma empresa concessionária de energia elétrica possa ser considerada confiável e de qualidade, é necessário que o sistema de transmissão garanta a entrega de energia aos pontos consumidores. No entanto, isso deve ser feito considerando que faltas em um sistema elétrico são inevitáveis, além de ser um problema do qual não há como fugir. Elas podem ocorrer de forma natural, por exemplo, um contato temporário devido a uma árvore ou simplesmente por pássaros, mitigação, descargas atmosféricas, rompimento dielétrico do ar por causa da poluição ou incêndios, entre outros (SAHA et al., 2012). As faltas também podem ocorrer de forma não natural, por exemplo: rompimento de condutores. Estas faltas ocorrem com bastante frequência nas linhas de transmissão de alta tensão devido à falta de isolamento em torno de cabos.

As linhas de transmissão são parte integrante da rede do sistema de energia elétrica e seu principal objetivo é transmitir a energia gerada ao consumidor com menos interrupções possíveis. Considerando que a demanda de energia elétrica cada vez é maior por causa do aumento da industrialização e urbanização do estilo de vida, uma análise de faltas rápida e precisa é essencial para um melhor desempenho e para minimizar as interrupções no sistema de distribuição de energia elétrica. Neste contexto, a detecção rápida e compensação de faltas em linhas de transmissão é muito importante para manter o normal funcionamento do sistema de energia elétrica.

Os sistemas de localização de faltas podem ser classificados em função de sua localização e das medidas elétricas que captam em um extremo da linha, ou nos dois extremos da linha. Os sistemas que utilizam valores medidos nos dois extremos da linha, devolvem resultados mais exatos que os sistemas que utilizam valores medidos em um único extremo da linha.

Em um sistema de potência podem-se apresentar vários tipos de faltas que causam perturbações, dentre as quais, as mais frequentes são as faltas fase-terra, presentes em 90% de eventos. Também, existem outros tipos de faltas, como são as fase-fase-terra, faltas fase-fase e as faltas trifásicas, todas elas com diferentes níveis de impedância. Na Figura 36 são ilustradas de forma resumida todos os tipos de faltas que podem acontecer nas linhas de transmissão.

Figura 36 - Tipos de faltas nas linhas de transmissão.



Fonte: Próprio Autor

Estes tipos de faltas podem-se apresentar em qualquer local da linha de transmissão e podem ser de baixa ou alta impedância. Geralmente as de alta impedância apresentam um valor muito alto no ponto de falta, o que causa uma corrente de baixa magnitude em comparação com as faltas sólidas (faltas de muito baixa impedância).

Uma característica muito importante associada às faltas de alta impedância é não ser linear. Uma falta deste tipo, esta geralmente associada a um arco elétrico que pode estar relacionado ao ponto de contacto do condutor com a terra ou outro objeto no tempo em que a falta esta acontecendo. Dadas estas características, as faltas de alta impedância podem não ser detectadas nos sinais de tensão e corrente utilizados pelos dispositivos de proteção e localização, enquanto uma baixa amplitude da corrente gerada pela falta de alta impedância pode ser confundida com um aumento de carga.

Quando uma falta acontece, geralmente, a subestação de energia elétrica que esteja mais perto, apresenta um comportamento anormal, tal como: aumento de corrente e queda de tensão nas fases que estão tendo o problema de falta. Este comportamento anormal depende da impedância na subestação, o que leva a mudanças nos fluxos de potência e o ângulo de transferência entre as duas subestações que estão interligadas, além de possíveis oscilações de frequência e de harmônicos de corrente e tensão.

Levando em consideração estas características, quando ocorre uma falta, dificilmente esta pode ser localizada. Como consequência, o tempo de reparo do defeito pode aumentar, acarretando em maiores custos operativos, diminuição da confiabilidade do sistema, prejuízo paras as grandes indústrias dado que seus processos de produção podem se ver interrompidos, além dos prejuízos à sociedade como um todo.

As faltas que acontecem nas linhas de transmissão se enquadram dentro de duas características que são faltas temporárias e faltas permanentes:

- As **faltas temporárias** são causadas normalmente por descargas atmosféricas e são corrigidas pelo religamento automático do sistema, porém sua correta localização é importante para a criação de um histórico confiável sobre as descargas atmosféricas ocorridas na região e estudos de desempenho e localização dos equipamentos de proteção do sistema de transmissão.
- As **faltas permanentes** são causadas pelo rompimento de condutores, isoladores com defeitos instalados em regiões com nível de poluição muito alto ou com incêndios, devido ao rompimento dielétrico do ar, queda de torres, entre outros, e a correção é feita pessoalmente por um funcionário da empresa de transmissão de energia.

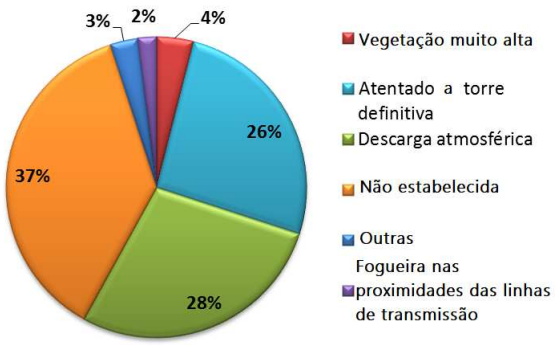
A localização rápida e eficiente da falta é de grande importância, dado que assim, a equipe habilitada para fazer a manutenção pode chegar ao local exato onde aconteceu a falta de forma rápida, por conseguinte o tempo total de reparo e de interrupção no fornecimento de energia é diminuído.

As principais causas de acontecimentos de faltas identificadas pelos analistas de perturbações são as seguintes:

- Vegetação muito alta.
- Atentados terroristas nas torres de energia elétrica.
- Descargas atmosféricas;
- Não estabelecidas.
- Incêndio nas proximidades das linhas de transmissão, acidental ou ocasionado, entre outras.

A localização de faltas tem que cobrir todas estas causas apresentadas na Figura 37, e determinar sua localização ao longo do seu percurso.

Figura 37 - Causas das faltas nos sistemas de transmissão de energia elétrica.



Fonte: Próprio Autor

Desta forma, é possível perceber a importância que tem o fato de realizar de forma rápida e correta a localização do local da ocorrência de uma falta, seja ela permanente ou temporária.

5.2 MÉTODOS UTILIZADOS PARA LOCALIZAÇÃO DE FALTAS

A pesquisa na área de localização de faltas tem aumentado nos últimos anos de forma gradativa. Este crescimento foi dado pela necessidade de minimizar os tempos de reparos dos sistemas de potência, permitindo aumentar tanto a confiabilidade quanto a qualidade do serviço, reduzindo assim, os prejuízos causados nas indústrias e na sociedade.

Atualmente podem ser encontrados na literatura especializada nesta área, vários estudos de métodos de localização de faltas. Cada nova pesquisa realizada permite apresentar novos métodos e aperfeiçoar os já existentes, com o objetivo de aumentar a precisão da localização do local no qual aconteceu a falta.

Segundo (KAWADY; STENZEL, 2002) existem alguns métodos de localização de faltas tais como:

- Métodos que têm como base a técnica de ondas trafegantes.
- Métodos que têm como base a técnica de medição de fasores em regime permanente.

Os métodos que têm como base a medição dos fasores em regime permanente, utilizam os dados registrados dos sinais de tensão e corrente. Isto faz com que sejam muito mais utilizados se comparados com os métodos que têm como base a técnica de ondas trafegantes, os quais necessitam de equipamentos específicos para que o método possa ser utilizado, o que torna a aplicação deste método mais caro e até inutilizável.

Os métodos de localização de faltas que têm por base a medição dos fasores em regime permanente são divididos em dois grupos:

- Métodos que utilizam dados apenas do terminal local.
- Métodos que utilizam dados de vários terminais.

Alguns exemplos de métodos de localização de faltas que têm sido desenvolvidos ao longo dos anos podem ser citados como:

Wang et al. (2008) desenvolveram um algoritmo imune à influência da capacitância shunt e à resistência de falta, baseado na precondição que o ângulo de fase da componente de sequência negativa da corrente medida em um dos terminais descreve precisamente o ângulo de fase da componente de sequência negativa da corrente de falta e da tensão de falta.

Em Lin et al. (2011) foi desenvolvido um método baseado em sistemas de redes neurais, utilizando várias configurações de falta para treinar a rede.

Em Faybisovich e Khoroshev (2008) apresenta-se um método que utiliza conteúdo espectral do transiente de tensão da alta frequência medido em dois terminais, fazendo desnecessária a sincronização dos sinais.

Gopalakrishnan et al. (2000) apresenta um algoritmo que utiliza o método das características para determinar as formas de ondas da tensão e da corrente e compará-las com as formas de onda medidas da tensão e da corrente.

Em Brahma e Girgis (2004) foi desenvolvido um método que se baseia nas variações de tensão causadas pela falta, utilizando somente os fasores de tensão medidos em ambos os terminais.

Liao e Kezunovic (2007) apresentaram um método baseado na teoria da estimativa não-linear para identificar os dados medidos como erro e estimar os referidos dados a fim de melhorar a precisão da localização de falta.

Fulczyk et al. (2008) desenvolveram um método de localização de faltas para linhas com circuitos duplos e compensação série que utiliza sub-rotinas para calcular a distância e a resistência de falta nos segmentos formados entre as barras e o banco de capacitores em série. Os resultados obtidos pelas sub-rotinas são analisados por um algoritmo para definir a saída válida dentre as saídas das sub-rotinas.

Em Liang, Wang e Li (2007) é apresentado um método que utiliza o algoritmo *Particle Swarm Optimization* em conjunto com o método dos mínimos quadrados, combinando a habilidade de convergência ótima global com a convergência rápida do Método dos Mínimos Quadrados.

Feng et al. (2008) desenvolveram um método baseado em ondas viajantes, onde são instalados analisadores que medem o tempo de chegada da onda, os quais são convertidos em tempo nas subestações de referência, onde ocorre a fusão com pesos de tempos convertidos.

Outro método proposto por Zhang et al. (2008) baseia-se nas ondas trafegantes. Neste método, medem-se ondas através de GPSs instalados em todas as subestações. O tempo de chegada da onda é definido e, os tempos medidos são utilizados para definir a distancia da falta eliminando a necessidade de se determinar a velocidade de propagação da onda.

Em Shahid et al. (2012) um método de classificação de faltas em redes inteligentes é desenvolvido utilizando algoritmos baseados nas MVS.

Sunil e Vishwakarma (2015) apresenta a aplicação de técnicas inteligentes utilizadas para diagnosticar faltas nas linhas de transmissão.

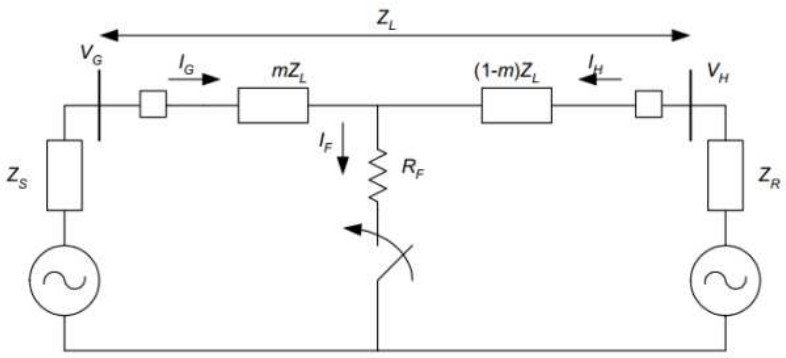
Em Wang, Aravinthan e Ding (2014) é apresentado um método que considera que as redes inteligentes requerem de sistemas de distribuição auto-curados com técnicas de detecção e classificação de falhas mais precisas. Um algoritmo de detecção e classificação de falhas de nível de alimentador multi-sensor é apresentado neste trabalho, com base nas técnicas da MVS. Foi utilizado um sistema de 34-barras com condições de carregamento dinâmico para avaliar o algoritmo desenvolvido. Foi aplicado o ruído nas medidas de corrente trifásica. A análise numérica indica que são alcançadas altas precisões na detecção e classificação de falhas para o algoritmo proposto.

O método desenvolvido por Liang e Wang (2008), aborda a localização de faltas em linhas de transmissão de circuitos duplos ou simples e com derivação. Este método baseia-se na medição dos fasores de tensão e corrente em todos os terminais da linha. Para a localização de faltas em linhas com vários terminais é aplicada a técnica de localização de faltas em redes de dois terminais. A fim de se determinar o correto local da falta e, por conseguinte a distância de ocorrência da falta, foi desenvolvido um procedimento através de índices de localização de faltas.

A técnica de localização de faltas em linhas com dois terminais utiliza as equações de propagação das ondas de tensão e corrente ao longo da linha para definir as tensões e correntes no ponto da falta, tendo como dados iniciais os fasores de tensão e corrente medidos em todos os terminais da linha.

Assim como os algoritmos clássicos mais utilizados na localização de faltas têm suas bases na equação de linha curta para linhas de transmissão, nesta pesquisa são utilizados os sinais de tensão e corrente obtidas a partir dos relés ou dos registradores de faltas e a informação da impedância da linha. Estes cálculos são afetados pelo aporte dos extremos da linha (efeito infeed) e pela resistência da falta (LIANG; WANG, 2008). A Figura 38 ilustra um diagrama básico para realizar esta análise.

Figura 38 - Esquema geral de linha curta.



Fonte: Adaptado de (LIANG; WANG, 2008)

Definições das variáveis do modelo de linha curta para linhas de transmissão de energia elétrica podem ser observadas a seguir:

- **m:** Localização da falta.
- Z_L**:** Impedância da sequencia positiva.
- R_f**:** Resistência da falta.
- I_f**:** Corrente da falta;
- I_a**:** Corrente da fase A;
- V_G**:** Tensão no nó G;
- V_H**:** Tensão no nó H;
- **Sub S:** Subestação S;
- **Sub R:** Subestação R;

A técnica de localização de faltas em linhas com dois terminais, utiliza as equações de propagação das ondas de tensão e corrente ao longo da linha para determinar as tensões e correntes no ponto da falta, tendo como dados iniciais os fasores de tensão e corrente medidos em todos os terminais da linha.

Com esta pesquisa pretende-se criar um novo método de localização de faltas de forma inteligente utilizando os SIs, especificamente, algoritmos de AM tais como: RNAs, MVS e ADs, e determinar sua precisão de localização mediante a comparação dos resultados gerados pelos sistemas de AM com os resultados obtidos em casos de faltas reais.

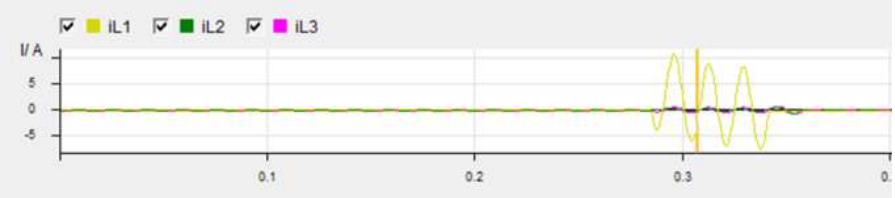
5.3 METODOLOGIA PROPOSTA

Nos STEE com frequência acontecem faltas nas linhas de transmissão que podem causar sua indisponibilidade. Nesta pesquisa é apresentada a aplicação de IC utilizada para determinar a localização de faltas nas linhas de transmissão dos STEE, realizando o treinamento dos SI com bancos de dados criados a partir das faltas simuladas e com localização de faltas reais que acontecem na linha de transmissão. Determinando a localização da falta de forma rápida e com um alto grau de precisão é possível diminuir o tempo de busca da localização física da falta por parte do pessoal de manutenção das empresas de transmissão de energia elétrica e realizar a normalização do serviço, minimizando os tempos de indisponibilidade. Esta tese, apresenta uma técnicas para detecção e classificação de faltas em linhas de transmissão de energia elétrica. A abordagem proposta baseiam-se no uso de IC, mais especificamente SIs.

Estes SIs são providos de algoritmos de AM tais como: em MVS, RNA e ADs, os quais realizam uma análise inteligente dos bancos de dados que contem o histórico de faltas reais criados a partir de monitoramento de diversos parâmetros elétricos ao longo da linha de transmissão.

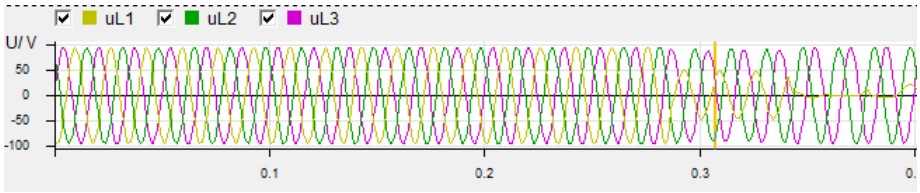
Alguns atributos considerados no banco de dados estão determinados pelo fato de que quando acontece uma falta nas linhas de transmissão de energia elétrica é criado um registro desta falta nos relés da subestação vizinha. Nas Figuras 39 e 40 podem ser visualizados os registros das correntes e das tensões exportadas dos relés. Note-se que os sinais dos registros estão totalmente filtrados pelos relés. De forma geral, para poder determinar a localização da falta deve ser utilizada a informação dos últimos ciclos.

Figura 39 - Sinal de corrente. Registro dos relés em uma perturbação real da falta.



Fonte: Próprio Autor

Figura 40 - Sinal de tensão. Registro dos relés em uma perturbação real da falta.



Fonte: Próprio Autor

A partir destes sinais de corrente e de tensão, se inicia a procura pela localização da falta, extraindo a seguinte informação, que será armazenada no banco de dados utilizado para o treinamento e validação dos SIs:

- Tensão de fases RMS.
- Correntes de fases RMS.
- Ângulo de tensão e corrente.
- Tensão e corrente.

Nesta pesquisa, para criar o banco de dados com o histórico de faltas no sistema de potência são utilizados, além dos dados reais, dados gerados a partir das simulações realizadas

no software Digsilent Power Factory. Este software permite simular as faltas quilométricas e gera dados tais como: tensão em p.u, correntes em kA, entre outros. A Tabela 21 ilustra alguns dados simulados pelo software Digsilent que foram utilizados na etapa de treinamento dos SI.

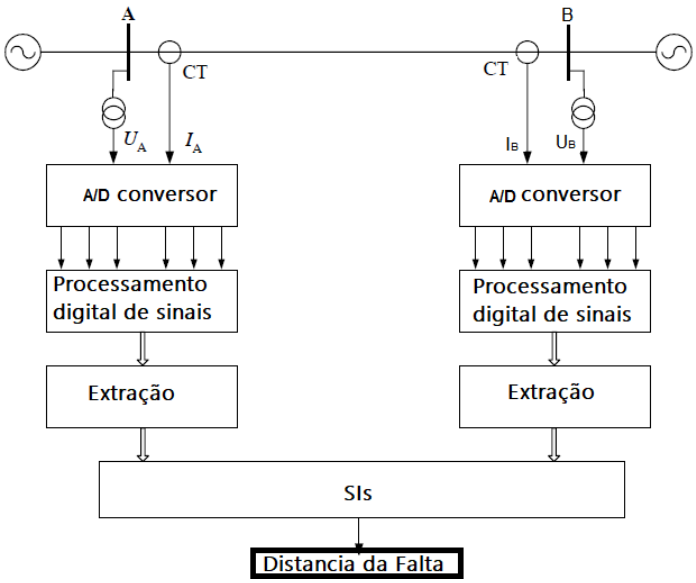
Tabela 21 - Dados gerados pelo Digsilent.

N°	V (p.u) Fase A	V (p.u) Fase B	V (p.u) Fase C	I (kA) Fase A	I (kA) Fase B	I (kA) Fase C	I (kA) Fase N
1	0,45	0,45	0,45	1,96	1,96	1,96	0,00
2	0,50	0,50	0,50	1,79	1,79	1,79	0,00
3	0,58	0,58	0,58	1,57	1,57	1,57	0,00

Fonte: Próprio Autor

O modelo de medição para obtenção de dados da linha de transmissão selecionada para a análise com a qual se definem as variáveis que são utilizadas no treinamento dos SI, é ilustrado na Figura 41.

Figura 41 - Esquema de medição utilizado para obtenção de dados.



Fonte: Próprio Autor

O sistema está equipado com duas fontes equivalentes de Thevenin as quais representam os aportes de falta nos extremos da linha. Na Tabela 22 podem ser analisados os valores de impedância da linha Primavera - Bacatá 500 kV do sistema de transmissão colombiano ilustrado na Figura 42. A linha de transmissão utilizada para realizar os testes desta pesquisa é a que aparece na Figura 42 em cor vermelha.

Tabela 22 - Parâmetros da linha de transmissão.

Parámetro	Primavera – Bacatá 500 kV
Longitude (km)	196.92
R1 (Ω)	4.59
X1 (Ω)	65.16
B1 (uS)	951.38
R0 (Ω)	75.61
X0 (Ω)	192.19
B0 (uS)	578.86

Fonte: (PARÂMETROS. . . , 2016)

Figura 42 - Diagrama unifilar da rede de transmissão do sistema Colombiano.

Fonte: (DIAGRAMA. . . , 2016)

O modelo para realizar a localização de faltas fazendo uso de SIs esta composto pelos seguintes elementos:

- **Tipo de falta:** Dado que nos sistemas de potência existem vários tipos de faltas, desde o início tem que ser definidos para que tipo de falta será treinado o SI.
- **Localização da falta:** Considerando que as faltas podem acontecer em qualquer local ao longo da linha, a ideia é considerar diferentes localizações dentro da análise.

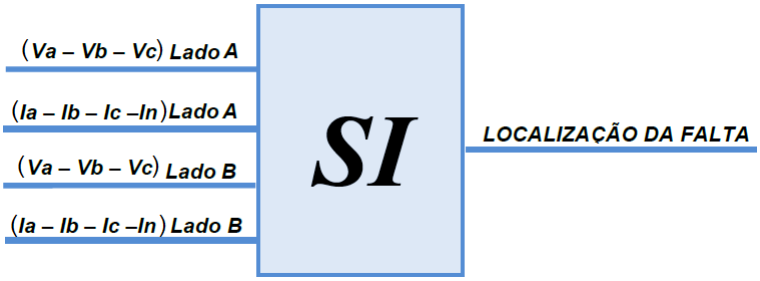
- **Potência de transferência:** São consideradas dentro da análise várias situações operativas de transferência.
- **Equivalência de cada extremo:** Se representam os valores das impedâncias equivalentes de cada um dos extremos.

Para a definição do banco de dados foi analisada a relação entre as variáveis de entrada e o resultado esperado. As variáveis escolhidas são as seguintes:

- Tensão das três fases (pu).
- Corrente das três fases (pu).
- Corrente de neutro (pu).
- O ângulo entre a tensão e a corrente (graus).

Para realizar o treinamento dos SIs é utilizada a informação dos dois extremos da linha de transmissão, como mostrado na Figura 43.

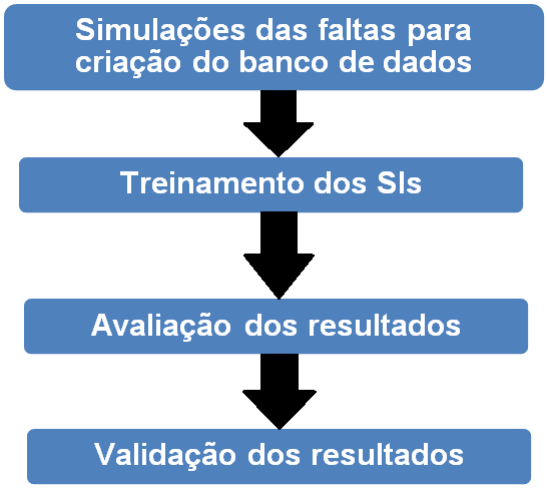
Figura 43 - Modelo dos SIs com suas variáveis de entrada e saída.



Fonte: Próprio Autor

Na Figura 44, é apresentado o processo passo a passo que deve ser realizado para treinar e avaliar os SIs, lembrando que o banco de dados está conformado por dados simulados e dados reais de faltas na linha de transmissão.

Figura 44 - Processo descritivo do funcionamento dos SIs.



Fonte: Próprio Autor

Depois de ter realizado os processos de treinamento e validação dos SI, é necessário determinar a precisão dos algoritmos de AM na tarefa de localização de faltas. Para tal, é realizado um procedimento que permite determinar a qualidade dos SIs mediante a definiçaõ do *erro*, o qual esta determinado pela seguinte expressão:

$$erro(\%) = \frac{d - d_{exact}}{L} \tag{30}$$

Onde:

- **d:** distância estimada da localização da falta em km ou em pu;
- d_{exact}**:** localização exata da falta em km ou em pu;
- **L:** Comprimento total da linha em km ou em pu

Considerando esta medida de erro, foi realizada uma comparação entre os métodos tradicionais e os novos métodos de SIs, entre eles as MVS, RNAs e ADs aplicados à localização de faltas, obtendo bons resultados.

5.4 CASO DE ESTUDO E RESULTADOS

Como estudo de caso foi analisado o problema de faltas em um STEE colombiano administrado por uma empresa de transmissão de energia elétrica (ETEE) colombiana, que pretende melhorar o processo de localização de faltas com o intuito de minimizar o tempo de correção das mesmas e assim garantir um melhor serviço para seus clientes.

As bases de dados utilizadas nas etapas de treinamento e validação dos SI, contém dados reais e simulados que foram fornecidos pela ETEE.

Para realizar o treinamento dos SIs são simuladas grandes quantidades de faltas na linha de transmissão fazendo uso do software Digsilent Power Factory, e foram realizadas variações das seguintes condições:

- O equivalente em cada um dos extremos A e B.
- Variação das condições de carga.
- Se realizam faltas quilométricas do 0%, até o 100% em passos de 5%.
- Se realizam faltas com diferentes valores de impedância, iniciando em 0 até chegar em 50 Ohm, usando passos de 5 Ohm.
- Simular os quatro tipos de faltas: trifásica, monofásica, bifásica e bifásica a terra.

Com os resultados das simulações é criado o banco de dados necessário para realizar o treinamento e validação dos SIs. Este banco de dados está conformado pelos dados de entrada reais e simulados, dentre os quais existem: 6 valores de tensão para três fases, 8 valores de correntes para as três fases e o neutro, e um atributo que contem os resultados da localização da falta. Portanto, esta base de dados contem um total de 15 atributos e 5544 cenários diferentes, ou seja, 83160 registros , os quais correspondem aos dados dos casos que foram analisados.

A base de dados utilizada para realizar o treinamento e validação dos SI está conformada por 15 atributos, dos quais 14 correspondem aos dados de entrada ao SI e o último corresponde à localização da falta para cada cenário ou também chamada de saída do SI. Os atributos selecionados para analisar este problema são:

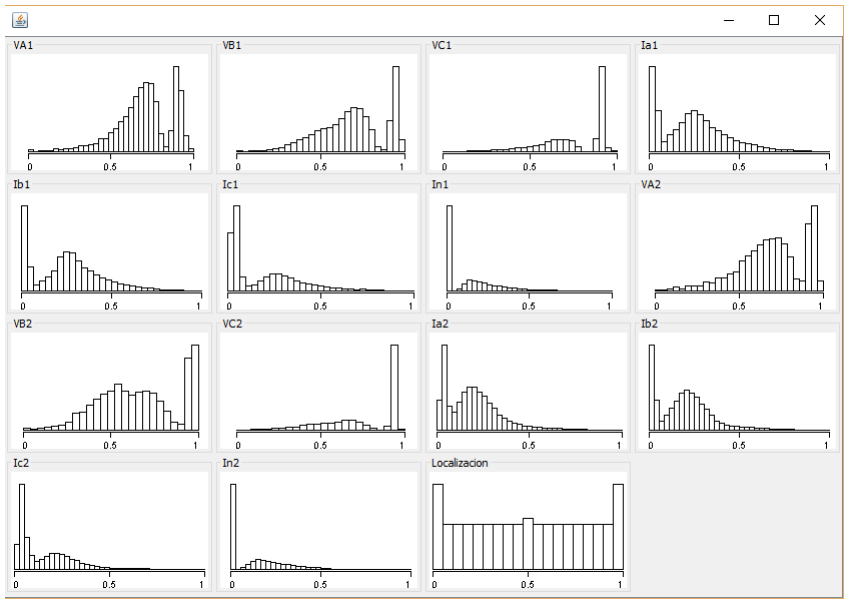
- **VA1:** Tensão da linha A1;
- **VB1:** Tensão da linha B1;
- **VC1:** Tensão da linha C1;
- **Ia1:** Corrente da fase a1;
- **Ib1:** Corrente da fase b1;
- **Ib1:** Corrente da fase c1;
- **In1:** Corrente do neutro n1;
- **VA2:** Tensão da linha A2;

- **VB2:** Tensão da linha B2;
- **VC2:** Tensão da linha C2;
- **Ia2:** Corrente da fase a2;
- **Ib2:** Corrente da fase b2;
- **Ib2:** Corrente da fase c2;
- **In2:** Corrente do neutro n2;
- **Localização:** Localização da falta;

A base de dados inicialmente fornecida pela ETEE passou por um processo de formatação dos dados para possibilitar o acesso com a ferramenta WEKA. Após este processo de formatação, o novo banco de dados é utilizado para realizar o treinamento e validação dos SIs. Neste caso são utilizadas as AD, as RNA e as MVS, e todos os resultados são avaliados com dados reais da empresa fornecedora da base de dados.

Uma vez que os dados são formatados, o passo a seguir é acessar o banco de dados fazendo uso da ferramenta WEKA. Após o carregamento dos dados, é aplicado um filtro de normalização dos dados. O resultado da normalização é ilustrado na Figura 45

Figura 45 - Dados filtrados.



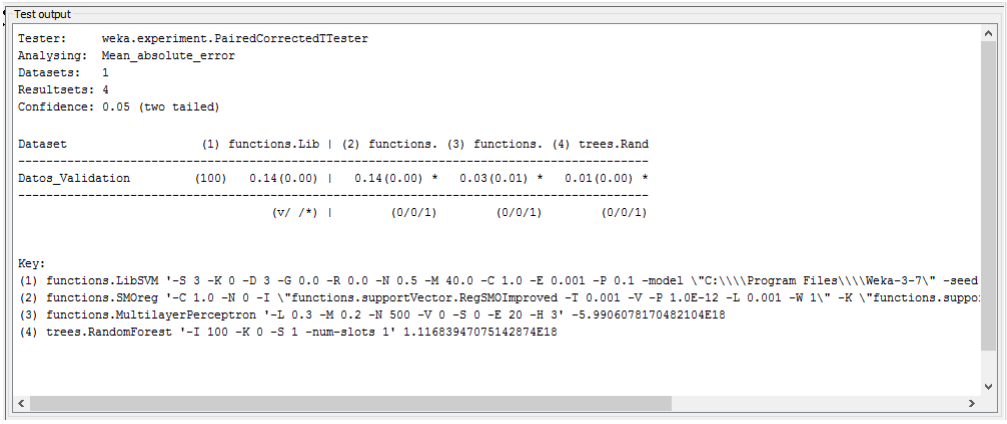
Fonte: Próprio Autor

Para avaliar a precisão dos SI devem ser consideradas medidas tais como:

- O erro médio absoluto *(Mean absolute error)*.
- A raiz do erro médio ao quadrado *(Root mean squared error)*.
- O erro absoluto relativo *(Relative absolute error)*.
- A raiz do erro relativo ao quadrado *(Root relative squared error)*.
- Desvio padrão *(Standard Deviations)*.

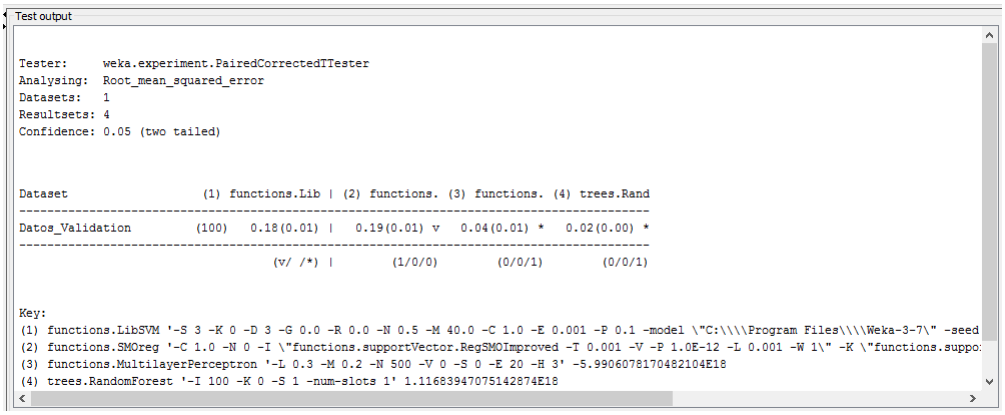
Nas Figuras 46, 47, 48 e 49 é ilustrado cada um destes items e o desvio padrão para cada um deles. Estes resultados foram obtidos fazendo a comparação entre dois algoritmos de MVS, um algoritmo de RNAs e um algoritmo de ADs.

Figura 46 - Comparação do erro médio absoluto.



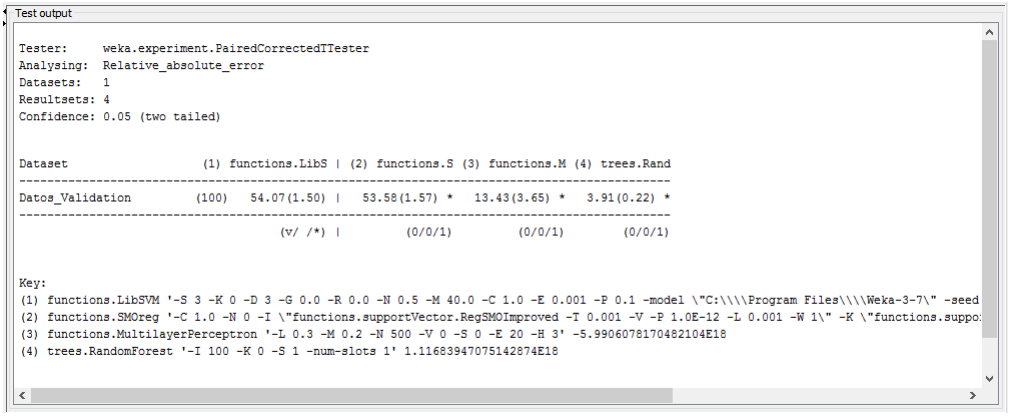
Fonte: Próprio Autor

Figura 47 - Comparação da raiz do erro médio quadrado.



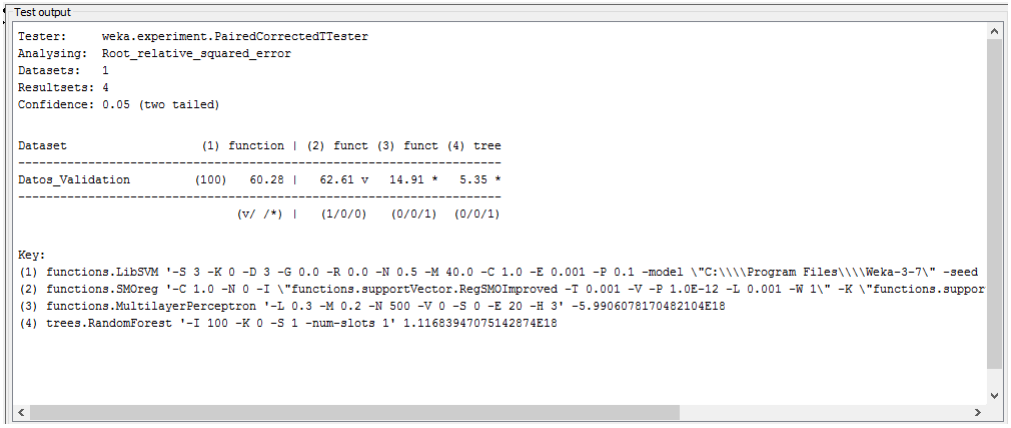
Fonte: Próprio Autor

Figura 48 - Comparação do erro absoluto relativo.



Fonte: Próprio Autor

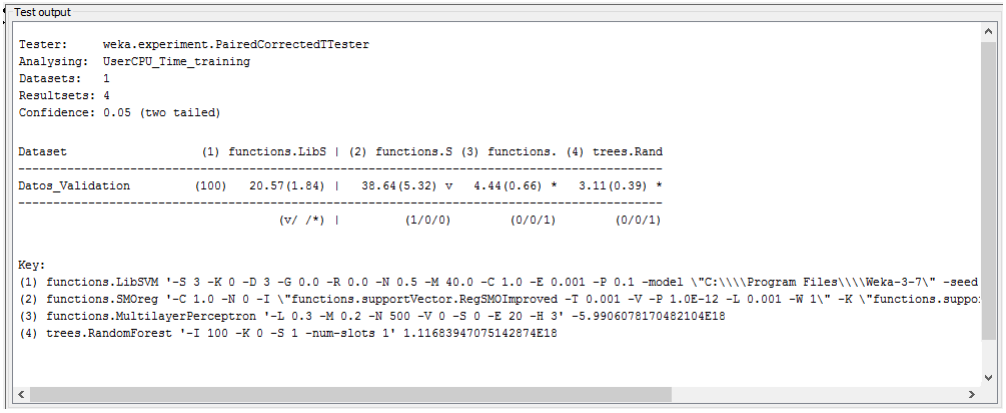
Figura 49 - Comparação da raiz do erro relativo ao quadrado.



Fonte: Próprio Autor

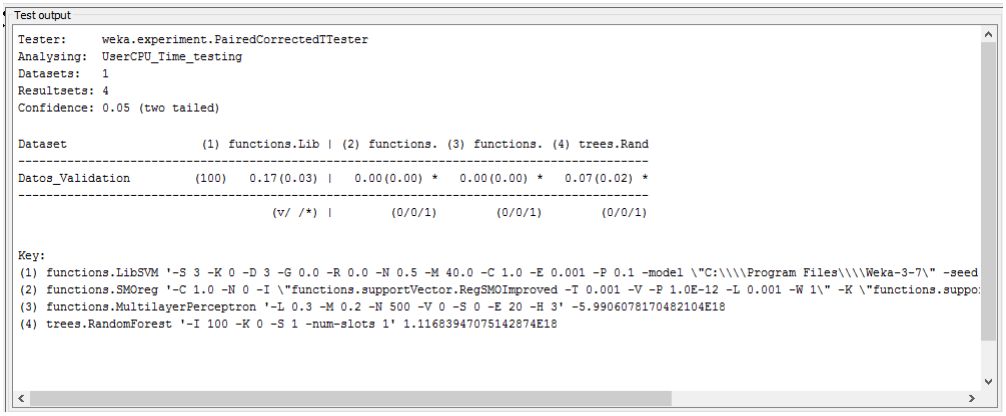
Considerando o tamanho do banco de dados, resulta fácil pensar que o tempo computacional para executar estes testes pode ser elevado, mas nas Figuras 50 e 51, são ilustradas as comparações entre os tempos computacionais para cada um dos algoritmos, tanto na parte de treinamento quanto na validação. Estes testes foram realizados em um computador core i7, com 8 GB de RAM, utilizando o sistema operacional DEVIAN 7.

Figura 50 - Comparação do tempo de treinamento dado em segundos.



Fonte: Próprio Autor

Figura 51 - Comparação do tempo de validação dado em segundos.



Fonte: Próprio Autor

Na Tabela 23 pode ser visualizada a porcentagem de erro que foi calculada considerando as impedâncias de falta de 0, 25 e 50 Ohm. Através das informações contidas nesta tabela pode ser análisado que, com o algoritmo de ADs a porcentagem de erro é menor que com o algoritmo de RNA e de MVS, chegando a um valor de 3,4625%, o que é considerado uma solução de qualidade, quando comparada à entregue pelos métodos de solução clássicos usados pela empresa em questão.

Tabela 23 - Medidas estatisticas do erro absoluto.

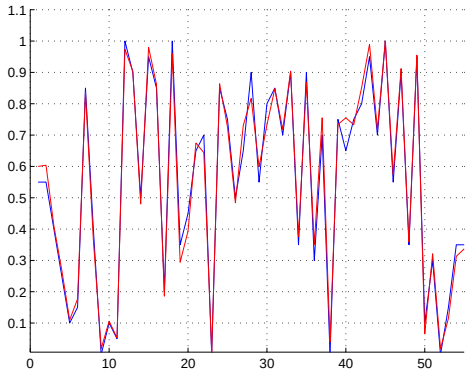
	MVS		AD	RNA
Erro (%)	10,0636	9,7081	3,4625	3,9064

Fonte: Próprio Autor

As Figuras 52-55 ilustram graficamente o comportamento do valor definido pelo SI (na cor vermelho) quando comparado com o valor real (na cor azul). Esta comparação foi realizada

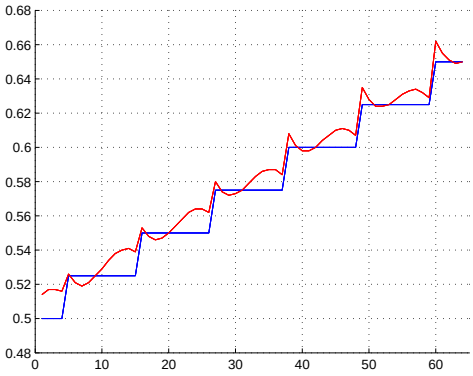
com os dois algoritmos que demonstraram melhores resultados.

Figura 52 - Localização de faltas ADs Vs falta simulada.



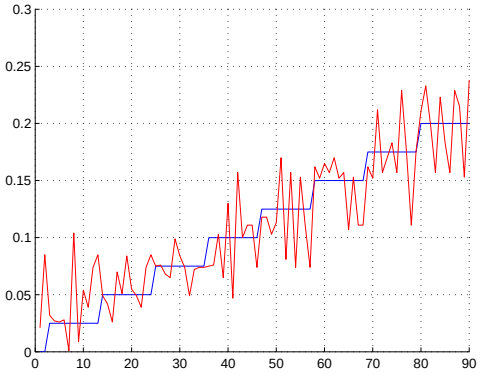
Fonte: Próprio Autor

Figura 53 - Localização de faltas RNA Vs falta simulada.



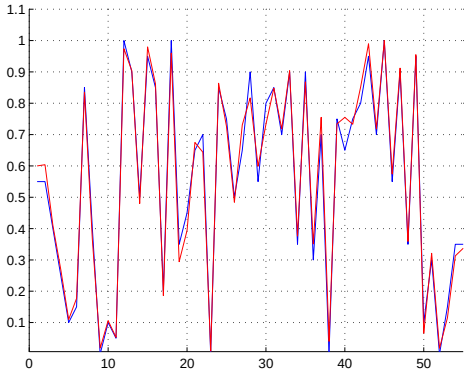
Fonte: Próprio Autor

Figura 54 - Localização de faltas RNA Vrs falta simulada.



Fonte: Próprio Autor

Figura 55 - Localização de faltas ADs Vrs falta simulada.



Fonte: Próprio Autor

5.5 CONCLUSÕES DO CAPÍTULO

A comparação entre os resultados obtidos na localização da falta de forma clássica realizada pela empresa fornecedora dos dados, e os resultados obtidos usando os SIs levam a considerar o uso destas metodologias como alternativas autónomas, rápidas e eficientes, na hora em que se busca a minimização do tempo de localização e ações de correção da falta.

Esta metodologia que permite a aplicação de SIs para realizar a localização de faltas nos STEE, demonstra ser eficiente e rápida na hora de realizar os testes e a validação. Isto quer dizer, que podem ser implementadas dentro do contexto das redes inteligentes com o intuito de resolver o problema de localização de faltas em tempo real.

Os métodos clássicos de localização de faltas são ferramentas práticas e entregam resultados aceitáveis, mesmo assim, utilizar SIs para realizar localização de faltas permite realizar o processo de forma rápida, eficiente e automatizada.

Usar SIs permite ter resultados mais exatos na localização de faltas e, garante a diminuição de forma significativa dos tempos de localização da falta por parte do pessoal de operação e das atividades logísticas para sua solução.

Com o uso de SIs previamente treinados, é possível desenvolver um método de localização de faltas capaz de localizar o trecho da linha onde acontece determinada falta em uma linha de transmissão.

Segundo os resultados obtidos, é possível afirmar que esta metodologia pode ser aplicada à localização de faltas em qualquer linha do STEE utilizando medições dos dois extremos da linha.

De acordo com os resultados apresentados pelos métodos baseados em SIs, pode-se afirmar

que, para as condições de teste consideradas, as ADs apresentaram um melhor desempenho com menores percentagens de erro e maiores percentagens de acerto. Os baixos percentagens de erro, tanto para as ADs, as MVSs, quanto para as RNAs, indicam que podem ser considerados como aceitáveis.

Os tempos de validação dos SIs de no máximo 0,17 segundo, são tempos aceitáveis quando comparados com os tempos de resposta de um controlador do sistema que realiza a localização da falta de forma manual.

6 SISTEMAS INTELIGENTES PARA COORDENAÇÃO DE CARGA ÓTIMA DE VEÍCULOS ELÉTRICOS E DISPOSITIVOS DE ARMAZENAMENTO CONSIDERANDO A TECNOLOGIA V2G

Neste capítulo é apresentada a aplicação da IC como solução ao problema de carregamento de veículos elétricos (VEs) e dispositivos de armazenamento (DAs). Este capítulo está organizado da seguinte forma: na Seção 6.1 é realizada uma introdução ao problema, e são apresentados os diferentes métodos de solução encontrados na literatura. Na Seção 6.2, é apresentada a metodologia de solução proposta. Na Seção 6.3 apresenta-se o caso de estudo e a análise dos resultados. Por último, Na Seção 6.4 apresentam-se as conclusões do capítulo.

6.1 INTRODUÇÃO

Com o passar dos anos, vários fatores tem gerado uma preocupação na área dos sistemas de transporte, alguns desses fatores são: aumento na demanda por mobilidade individual e grupal, emissões graves de poluentes que levam ao aquecimento global, entre outros. Segundo IEA (2017), só no ano de 2012 o setor de transporte a nível mundial consumiu 63,7% do petróleo o que levou a emissão de 7135 milhões de toneladas de dióxido de carbono para o meio ambiente. Estas cifras realmente são alarmantes considerando as projeções de mudanças climáticas.

Além da crise ambiental associada ao grande uso de combustíveis fósseis, existe outra preocupação que tem incentivado a eletrificação do transporte e o uso extensivo de VEs, esta é, a escassez de combustíveis provenientes de fósseis. Considerando estas problemáticas, Cao et al. (2016) e Cao et al. (2017) afirma-se que os VEs são as alternativas eficientes e sustentáveis mais indicadas para os sistemas de transporte privados e públicos. Além disso, Yilmaz e Krein (2013) e Hu et al. (2015) garantem que os VEs são uma das tecnologias mais promissoras, devido às suas capacidades de armazenamento de energia e potenciais interações com redes de energia renováveis.

Os governos de vários países, entre eles, Austrália, Canadá, China, União Européia e Estados Unidos, têm incentivado à população a comprar VEs através da implementação de várias ações, tais como: isenção de impostos, subsídios ou financiamento para compra deste tipo de veículos, construções das instalações de estacionamento dos VEs , entre outras, (VILLALOBOS et al., 2014).

Embora o incremento da implementação de VEs possa fornecer uma solução para as preocupações a nível mundial de falta de combustível fóssil e poluição do ar, tem que ser considerado que o aumento dos VEs ligados da rede elétrica também traz múltiplos problemas técnicos

para os SDEE, tais como desvios de tensão, sobrecarga de transformadores e do SDEE, incremento de perdas elétricas, etc, comprometendo a segurança e a confiabilidade da rede elétrica (WANG; XIAO; WANG, 2017) e (SABILLÓN et al., 2015). Como consequência, os SI como ferramenta de coordenação do carregamento inteligente dos VEs se tornam um importante tema de pesquisa.

No entanto, antes de pensar em SIs utilizados para a coordenação de carga inteligente dos VEs (CCIVEs), deve ser projetado um mecanismo de carregamento de VEs flexível e eficiente para coordenar dinamicamente o carregamento de VEs e satisfazer os requisitos dos SDEE. Só assim é possível treinar um sistema inteligente para realizar a CCIVE de forma automatizada. Sendo assim, é necessário resolver algumas questões técnicas, tanto dos SDEE quanto dos VEs.

Neste contexto, cabe lembrar que atualmente as baterias de tração são o suporte de energia elétrica mais comum, portanto, seu papel dentro dos sistemas de transporte é de suma importância quando é considerado o desempenho, a economia e aceitação dos VEs, (JAFARI et al., 2015) e (TANG; RIZZONI; ONORI, 2015).

Portanto, uma das preocupações que surgem é como tornar as baterias seguras, eficientes e duráveis no ambiente complexo do VE, para isto, é necessário monitorar e controlar cuidadosamente os estados internos da bateria, por exemplo, o estado de carga (EoC). A incerteza do EoC pode frustrar o roteamento do veículo e exacerbar os problemas de segurança e durabilidade da bateria (MENG; LUO; GAO, 2016). Para contornar este problema, varias pesquisas tem sido desenvolvidas para estimar o EoC da bateria, tais como as apresentadas em Kim (2008) e Kim (2010), cada um destes trabalhos apresenta suas vantagens e desvantagens.

Considerando a atual necessidade de realizar um ótimo controle e coordenação de carga para as baterias dos VEs em um determinado período de tempo, surgem pesquisas que estudam o problema da coordenação de carga dos VEs, visando, além de acelerar a carga da bateria, a programação de carga ótima que deve maximizar a energia carregada para as baterias dos VEs e minimizar as perdas de energia. Além disso, deve ser definida uma operação econômica para o SDEE, que satisfaça as restrições operacionais, (SABILLÓN et al., 2016).

Pesquisas encontradas na literatura especializada, abordaram o problema de coordenação de carga dos VEs em SDEEs. Uma destas pesquisas é apresentada em Trippe, Massier e Hamacher (2013), onde pretende-se realizar um carregamento de VEs de forma inteligente, considerando que os custos de carregamento são otimizados usando um modelo de MILP. Nessa pesquisa são considerados os perfis de carga da bateria de forma detalhada, as exigências de mobilidade e os limites de potência máxima do VE, mesmo assim, o impacto da carga de VEs na rede de distribuição de energia elétrica não é considerado.

Em O'Connell, Flynn e Keane (2014) foi apresentada uma solução em tempo real para a coordenação de carga dos VEs nos SDEE. Esta pesquisa apresenta um modelo de programação

não-linear que pretende otimizar o custo da carga de VEs, ignorando o custo de energia das cargas no SDEE. A implementação de esquemas de controle de carga, pode exigir previsões de uma série de variáveis, por exemplo, cargas domésticas, disponibilidade do VE e requisitos de bateria. A imprevisibilidade do comportamento individual do cliente pode, no entanto, levar a grandes variações entre os comportamentos previstos e realizados.

Levando em consideração o dito até agora e o relatado em Bilgin et al. (2015), onde se evidencia que tanto a indústria automotiva quanto a comunidade acadêmica têm-se dedicado a promover conceitos de mobilidade de alta eficiência, limpa e acessível, como uma forma de combater o aquecimento global, surge a necessidade de utilizar algoritmos avançados de IC que permitam realizar uma tomada de decisões de forma rápida e eficiente.

Neste contexto, os SIs podem ser aplicados com o intuito de gerenciar um sistema de CCI-VEs levando a uma operação do SDEE adequada.

Os algoritmos que compõem os SIs utilizados no desenvolvimento desta pesquisa são utilizados para realizar a coordenação de carga ótima de forma inteligente dos VEs e dos DAs nos SDEE. O SI define uma programação de carregamento ótimo durante um dia para VEs e DAs . Para tal, são consideradas informações e medições disponíveis no centro de controle, tais como: horários de chegada e de saída dos VEs, e seu estado inicial de carga. A energia inicial, a energia esperada no final do período de tempo e também são considerados os limites de descarga permitida para os DAs. O objetivo é minimizar o custo da operação dos SDEE considerando o carregamento das baterias dos VEs e dos DAs tentando atingir o nível mínimo de energia esperado para um período de tempo específico.

Os algoritmos que compõem os SIs podem ser treinados a partir das medições disponíveis e ações registradas no centro de controle do sistema. Entretanto, é usado um modelo matemático de programação linear inteiro misto (MPLIM) proposto em Sabillón et al. (2015) para criar este conjunto de medições e ações. O modelo MPLIM foi implementado na linguagem de modelagem matemática AMPL Fourer, Gay e Kernighan (2002) e foi resolvido usando o solver comercial CPLEX CPLEX (2009). Para demonstrar a eficiência da técnica de solução proposta, vários testes foram realizados usando um sistema de distribuição de 34 nós.

6.1.1 Métodos utilizados para coordenação de carga de Veículos Elétricos

Várias pesquisas encontradas na literatura especializada, abordaram o problema de coordenação de carga dos VEs em SDEEs. Uma destas pesquisas é (TRIPPE; MASSIER; HAMACHER, 2013), que pretende realizar um carregamento de VEs de forma inteligente, onde os custos de carregamento são otimizados usando um modelo de MILP, considerando: os perfis de carga da bateria de forma detalhada, as exigências de mobilidade e os limites de potência máxima, mesmo assim, o impacto da carga dos VEs na rede não é considerado.

Há outras pesquisas na literatura que levam em consideração a tecnologia veículo-a-rede (Vehicle-to-grid V2G) que permite coordenar o carregamento dos VEs.

Por exemplo, em Tang, Zhong e Bollen (2016) é proposto um novo método de modelagem de VEs e uma estratégia de controle de carga V2G ideal para grande quantidade de VEs. As estratégias de controle de carga de VEs e tecnologias V2G podem utilizar as propriedades dos VEs para obter vários benefícios. Os VEs são modelados de forma individual em algoritmos de controle de carga existentes. Um novo método de modelagem de VEs e uma estratégia de controle de carga ótimo são propostos para que todos os VEs, em uma área de controle, possam ser considerados como um único objeto no processo de otimização. A estratégia minimiza o custo de carga total dos VEs e pode ser expandida para atender o controle utilizando a tecnologia V2G. Com o novo método de modelagem, a carga computacional do algoritmo de otimização pode ser significativamente reduzida e não aumenta com o número de VEs. Assim, a estratégia é extremamente eficaz quando o número de VEs se torna grande e o custo de implementação pode ser mais razoável, pois é necessária menos capacidade computacional.

Em Sabillón et al. (2016), uma metodologia baseada em uma formulação de programação linear inteira mista é apresentada para resolver a coordenação de carregamento ótimo dos VEs em SDEE considerando a tecnologia V2G. A operação em estado estacionário do SDEE é representada usando as partes real e imaginária das magnitudes de tensão e correntes nos nós e nas linhas, respectivamente. A geração distribuída e o desequilíbrio dos circuitos e cargas do sistema são levados em consideração. O método desenvolvido define uma programação de carga ótima para os VEs. Este cronograma de carregamento considera os horários de chegada e partida dos VEs e seu estado de chegada, além da contribuição energética dos VEs equipados com tecnologia V2G.

Em Wang, Xiao e Wang (2017), é apresentado um esquema de controle de carga descentralizado - centralizado híbrido, que inclui três partes. Primeiro, uma abordagem de programação off-line é apresentada pela primeira vez, no lado de carregamento centralizado, com o objetivo de minimizar o custo da energia, ao mesmo tempo em que satisfaz os requisitos de carregamento dos VEs. Segundo, uma estratégia de programação adaptativa baseada em controle preditivo é desenvolvida para determinar os perfis de carregamento dos VEs, ótimos em tempo real, para lidar com a dinâmica e as incertezas do sistema. Terceiro, no lado do carregamento descentralizado, as interações entre VEs e o controlador do sistema de carregamento são modeladas como um jogo Stackelberg não cooperativo -líder-seguidor- em que o controlador do sistema atua como o líder e os VEs atuam como seguidores.

Em Shaaban et al. (2014), é proposto um novo método de coordenação on-line para o carregamento de veículos elétricos plug-in (VEPs) em redes de distribuição inteligentes. O objetivo do método proposto é carregar os VEPs de forma otimizada e minimizar os custos operacionais do sistema sem violar as restrições do sistema de energia. Ao contrário das soluções relatadas

na literatura, a arquitetura de carga proposta garante a viabilidade das decisões de carga por meio de uma nova unidade de previsão que pode prever a futura demanda de energia de VEPs e através de uma inovadora unidade de otimização de dois estágios que garanta uma coordenação de carga eficaz. A descarga coordenada do PEV também permite uma melhor utilização dos recursos do sistema de energia elétrica.

Outra pesquisa que relata uma proposta de solução ao problema de coordenação de carga de VEs é apresentada em Deilami et al. (2011), onde se propõe uma nova solução de gerenciamento de coordenação de carga para múltiplos VEPs em uma rede inteligente. Um algoritmo baseado em sensibilidades é proposto para a coordenação de carga de VEs em tempo real considerando as chegadas e partidas aleatórias dos VEs, o perfil de tensão e limites de geração de energia, a fim de minimizar o custo total de energia. A abordagem reduz o custo de geração incorporando preços de energia no mercado variável e considerando os fusos horários de tarifação preferencial do proprietário do PVE com base na seleção de prioridade. Esta abordagem permite que os VEPs sejam carregados o mais rápido possível, ao mesmo tempo que cumprem os critérios de operação da rede.

Na literatura especializada, também, encontram-se métodos de solução ao problema de coordenação de carga de VEs utilizando agentes inteligentes. Em Karfopoulos e Hatziargyriou (2013) é desenvolvido um sistema multi-agente que permite realizar o controle de carga de VEs, baseado no Princípio de Equivalência de Certeza de Nash que considera os impactos da rede. Os autores consideram que o carregamento descontrolado de VEs pode afetar adversamente o funcionamento normal do SDEE

Em Mets, D'hulst e Develder (2012), são apresentadas duas abordagens diferentes utilizadas para agendar o carregamento de VEPs minimizando o pico de carga máxima que eles possam causar. Uma das abordagens é de otimização clássica usando programação quadrática, e uma segunda abordagem é baseada na análise de mercado para realizar a coordenação de carga, para tal, são utilizados sistemas multi-agentes que usam lances em um mercado virtual para atingir um preço de equilíbrio que corresponde à demanda e ao fornecimento.

Neste trabalho é apresentado um sistema de controle baseado em agentes que coordena o carregamento da bateria de VEs em SDEE. O objetivo do sistema de controle é carregar os VEs em horas em que se apresentam baixos preços da eletricidade considerando as restrições técnicas da rede de distribuição de energia elétrica. A tomada de decisão dos agentes é realizada mediante o uso de técnicas de busca e redes neurais. Testes onde pode ser vista a capacidade do sistema de controle para funcionar com sucesso quando o SDEE opera dentro de seus limites de carga e quando os limites de carga são violados apresentam-se. Este método, não só apresenta um esquema de tarifação ideal para operação normal, mas também um método de planejamento de emergência empregado quando as condições normais são restauradas após uma violação de limite técnico. No entanto, os VEs são desconectados para restaurar o desempenho adequado

do SDEE e evitar violações de limites técnicos.

Neste contexto, pode ser visto que a IC pode ser aplicada com o intuito de gerenciar um sistema de controle que permita realizar a coordenação de carga dos VEs e dos DA em SDEEs adequadamente.

6.2 METODOLOGIA PROPOSTA

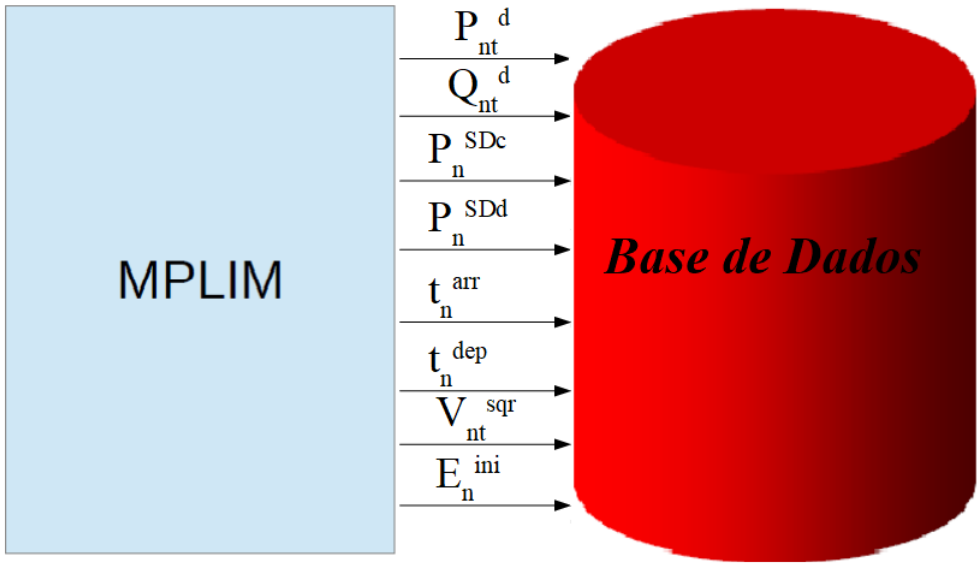
Nesta tese é apresentada uma metodologia baseada em SIs, que permite determinar a coordenação de carga para VEs e DA nos SDEE em tempo real, usando medições elétricas como base de dados de entrada, estes dados são submetidos a um processo de pré-processamento para posteriormente ingressar ao sistema inteligente. Dentro do SI é realizado o processo de classificação e finalmente se obtém a definição das variáveis que definem a coordenação de carga ótima para DA e VEs durante um período de tempo de um dia. A contribuição de energia dos VEs equipados com tecnologia V2G também é considerada.

Para o desenvolvimento desta pesquisa foi utilizado um MPLIM apresentado em Sabillón et al. (2015), onde são considerados os horários de chegada e saída dos VEs e seu estado inicial de carga. Além disso, é considerada a energia inicial, a energia esperada no final do período do dia e o limite de descarga permitida para os DA. Nesse trabalho foram consideradas as seguintes informações:

- As baterias dos VEs devem ser carregadas durante o tempo de chegada e saída do EV.
- O EoC inicial e o tempo de saída de cada EV é conhecido na hora da chegada.
- O operador pode definir o nível de energia esperado para cada DA no final de cada período.
- Os DAs não podem ser descarregados além do nível de descarga especificado.
- As baterias dos VE e os DA podem ser controladas pelo operador do SDEE através de dispositivos de comunicação. Este controle pode alterar o estado de carga de cada bateria e dos DA em cada etapa do tempo.

Este modelo de PLIM foi utilizado para criar as bases de dados necessárias para realizar o treinamento e validação dos algoritmos dos SIs. Na Figura 56 apresenta-se um diagrama que ilustra as informações que são abstraídas do MPLIM para conformar a base de dados utilizada nos processos de treinamento e validação dos SIs.

Figura 56 - Diagrama de abstração de dados do MPLIM para criação da base de dados.



Fonte: Próprio Autor

Para descrever o funcionamento dos SIs, é necessário conhecer os dados de entrada e de saída abstraídos do MLIM apresentado em Sabillón et al. (2015). Como dados de entrada do SI têm-se:

- $P^{d}_{n,t}$: Valores da potência ativa demanda na barra n na hora t;
- $Q^{d}_{n,t}$: Valores da potência reativa demanda na barra n na hora t;
- P^{SDc}_{n}: Valores da potência ativa demanda pelo DA na barra n;
- P^{SDd}_{n}: Valores da potência ativa injetada pelo DA na barra n;
- t^{arr}_{n}: Hora de chegada do VE na barra n;
- t^{dep}_{n}: Hora de saída do VE da barra n;
- $V^{sqr}_{n,t}$: Valores da tensão ao quadrado na barra n na hora t;
- E^{ini}_{n}: Estado de carga do VE na hora de chegada na barra n;

Como dados de saída do SI, têm-se o estado de carga dos VE e dos DA:

- $W_{n,t}$: Variável binária associada à carga e descarga da bateria do VE na barra n durante o intervalo de tempo t. Seu valor é 1 se a bateria está carregada no seu nível máximo;

- $X_{n,t}$: Variável binária associada à carga e descarga da bateria do VE na barra n durante o intervalo de tempo t. Seu valor é 1 se a bateria de um VE com tecnologia V2G esta injetando energia à rede no seu nível máximo;

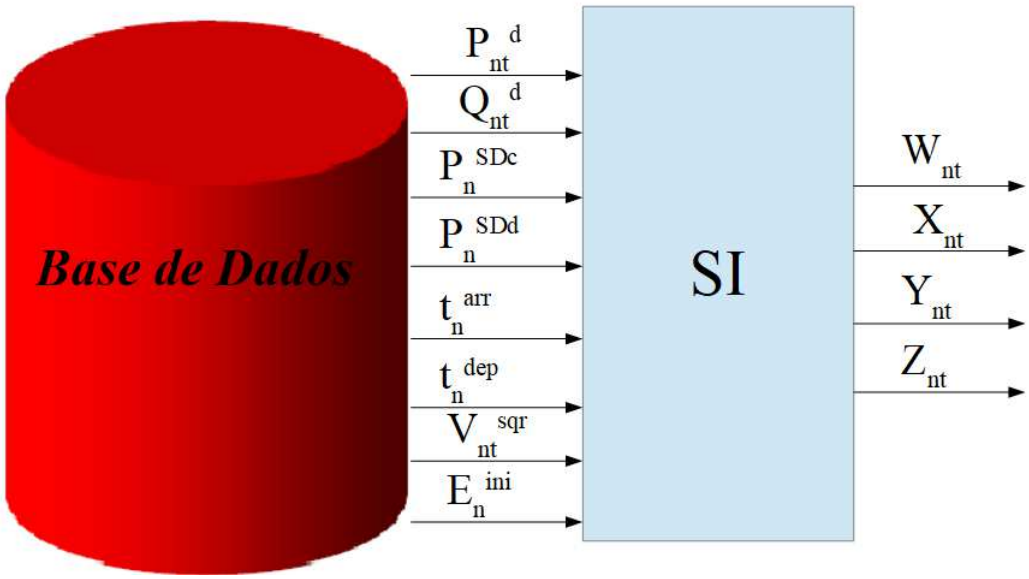
- $Y_{n,t}$: Variável binária associada à carga e descarga dos DA na barra n durante o intervalo de tempo t. Seu valor é 1 se o DA está carregada no seu nível máximo;

- $Z_{n,t}$: Variável binária associada à carga e descarga dos DA na barra n durante o intervalo de tempo t; Seu valor é 1 se o DA esta descarregado no seu nível máximo;

Uma vez criada a base de dados com medidas abstraídas a partir da execução do MPLIM, é realizado o processo de treinamento e validação dos algoritmos que compõem os SIs. Estes processos são realizados para cada algoritmo de forma separada. Na etapa de treinamento é ensinado ao algoritmo do SI como deve reagir a partir de determinados pares de dados de entrada e saída, isto é, o SI aprende a tomar decisões a partir de um conjunto de dados. Na etapa de validação são utilizados os modelos resultados da etapa de treinamento de cada algoritmo. Estes modelos que já foram previamente treinados com pares de dados de entrada e saída e são avaliados com novos dados que não contem as saídas. Esta etapa é realizada para medir o nível de aprendizado de cada algoritmo do SI. Este processo é repetido até encontrar o modelo de SI que apresente o melhor resultado.

Na Figura 57 ilustra-se um diagrama que mostra as respetivas variáveis de entradas e saídas do SI a partir da base de dados.

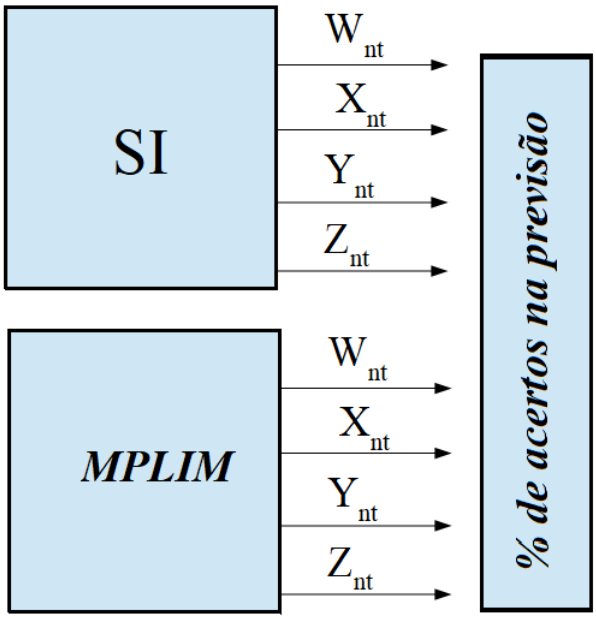
Figura 57 - Diagrama das variáveis de entradas e saídas do SI apartir da base de dados.



Fonte: Próprio Autor

Depois de realizar o treinamento e validação dos algoritmos que compõem os SIs, é realizado o processo de avaliação dos resultados obtidos para cada algoritmo de forma separada. Na Figura 58 ilustra-se um diagrama que mostra as respetivas variáveis de saídas do SI e do MPLIM. Estes valores contidos nestas variáveis são comparados para ver o índice de aprendizado dos SIs a partir da base de dados.

Figura 58 - Diagrama de avaliação de cada SI.



Fonte: Próprio Autor

Cada etapa do processo de treinamento e validação foram realizadas até obter o maior percentual de aprendizado.

6.3 CASO DE ESTUDO E RESULTADOS

Com o objetivo de treinar e validar os SIs propostos foram realizados testes com bases de dados com medições do sistema criadas a partir do MPLIM. Nesta pesquisa foi realizada uma análise do sistema durante 1 dia, hora a hora, para cada barra do sistema, portanto, foram obtidos em total 10.776 cenários, o que quer dizer que estão sendo analisados 161.655 dados. Estes cenários foram usados para realizar o treinamento dos SIs.

Uma vez tendo os sistemas treinados, foi criado outro banco de dados com cenários de 1 dia e mais 20% de novos cenarios sem valores de saída, hora a hora, para realizar a avaliação dos SIs. Para a aplicação desta metodologia foi utilizada a plataforma de aprendizado de máquina WEKA. Segundo os testes realizados com os SIs, os dados de treinamento apresentam um alto nível de veracidade na sua avaliação. Isto pode ser conferido nas figuras e tabelas com

os resultados obtidos que serão apresentadas nesta seção.

Os atributos que foram escolhidos para realizar o treinamento dos SIs são:

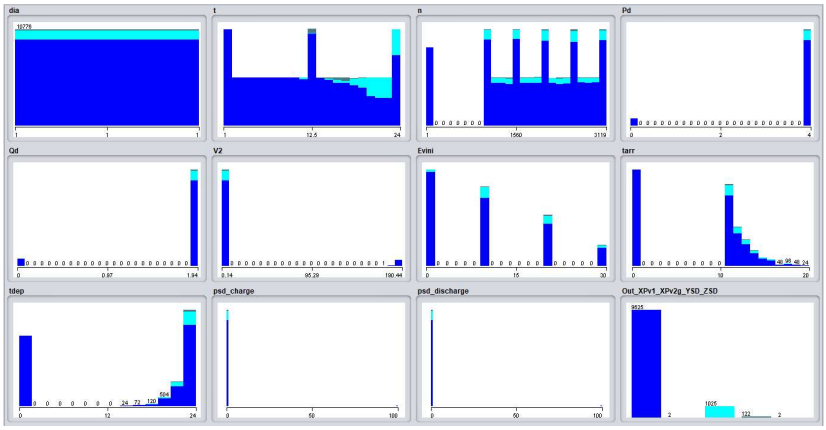
- DIA;
- HORA;
- BARRA;
- $P^{d}_{n,t}$: Valores da potência ativa demanda na barra n na hora t;
- $Q^{d}_{n,t}$: Valores da potência reativa demanda na barra n na hora t;
- P^{SDc}_{n}: Valores da potência ativa demanda pelo DA na barra n;
- P^{SDd}_{n}: Valores da potência ativa injetada pelo DA na barra n;
- t^{arr}_{n}: Hora de chegada do VE na barra n;
- t^{dep}_{n}: Hora de saída do VE da barra n;
- $V^{sqr}_{n,t}$: Valores da tensão ao quadrado na barra n na hora t;
- E^{ini}_{n}: Estado de carga do VE na hora de chegada na barra n;
- $W_{n,t}$: Variável binária associada à carga e descarga da bateria do VE na barra n durante o intervalo de tempo t;
- $X_{n,t}$: Variável binária associada à carga e descarga da bateria do VE na barra n durante o intervalo de tempo t;
- $Y_{n,t}$: Variável binária associada à carga e descarga dos DA na barra n durante o intervalo de tempo t;
- $Z_{n,t}$: Variável binárias associadas à carga e descarga dos DA na barra n durante o intervalo de tempo t;

Para uma melhor compreensão de cada processo que compõe as diferentes etapas de aprendizado dos SIs, são apresentados o processo de treinamento e de validação separadamente.

6.3.1 Treinamento e validação dos SIs

Os SIs são treinados e avaliados separadamente usando diferentes bancos de dados. A forma de treinar os SIs neste caso de estudo é realizada da seguinte maneira: inicialmente MPLIM entrega os valores calculados para cada uma das variáveis de controle de carregamento dos VEs e dos DA. Estes valores são utilizados para criar uma base de dados. A Figura 59, ilustra os dados que compõem a base de dados no seu estado original.

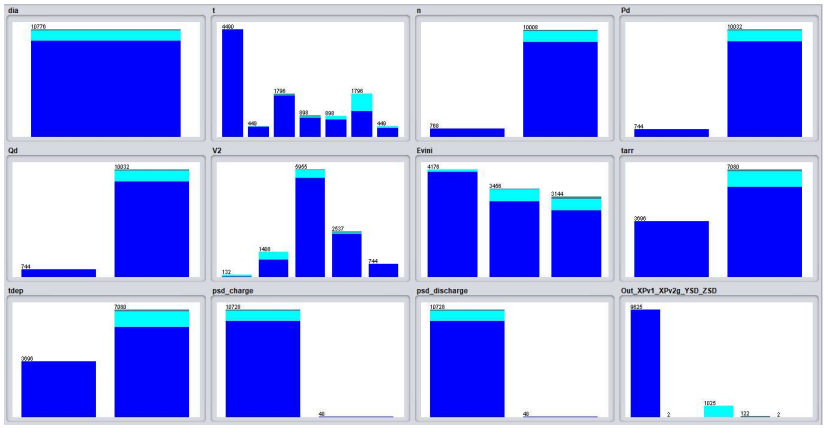
Figura 59 - Dados que compõem a base de dados inicial.



Fonte: Próprio Autor

Uma vez que a base de dados está pronta, é aplicado um filtro de discretização aos dados para realizar uma limpeza à base de dados que é utilizada para realizar a etapa de treinamento dos algoritmos que compõem os SIs. A Figura 60, ilustra os dados que compõem a base de dados depois de aplicar o filtro de discretização.

Figura 60 - Dados da base de dados apos à aplicação do filtro.

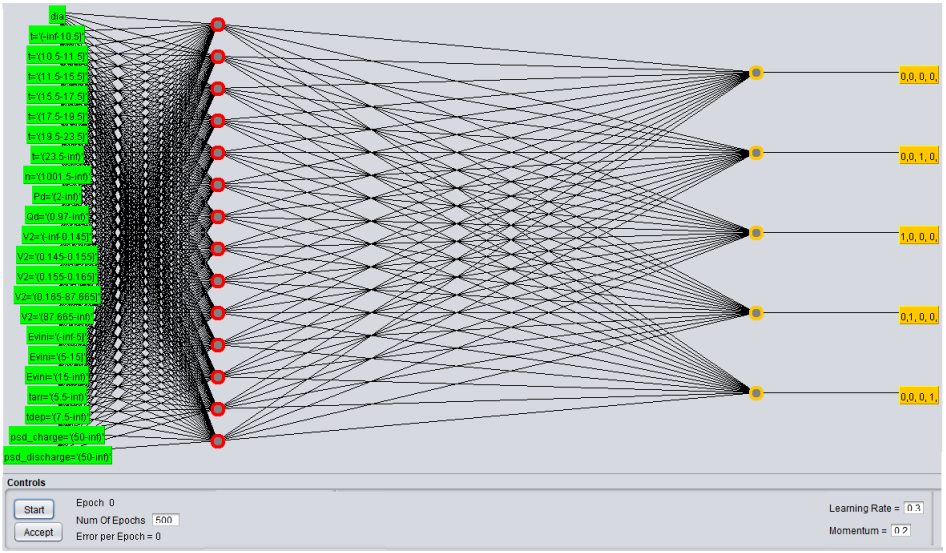


Fonte: Próprio Autor

O passo a seguir é treinar de forma separada os algoritmos que compõem os SIs. No

caso das RNA foram testadas varias arquiteturas até chegar na mais adequada para o problema em questão. Na Figura 61 é ilustrada uma RNA MLP treinada utilizando o algoritmo Back-propagation onde foram realizadas 500 iterações e as camadas ocultas são definidas de forma automatica utilizando a seguinte função $HL = (atributos + classes)/2$ definida em Witten et al. (2016).

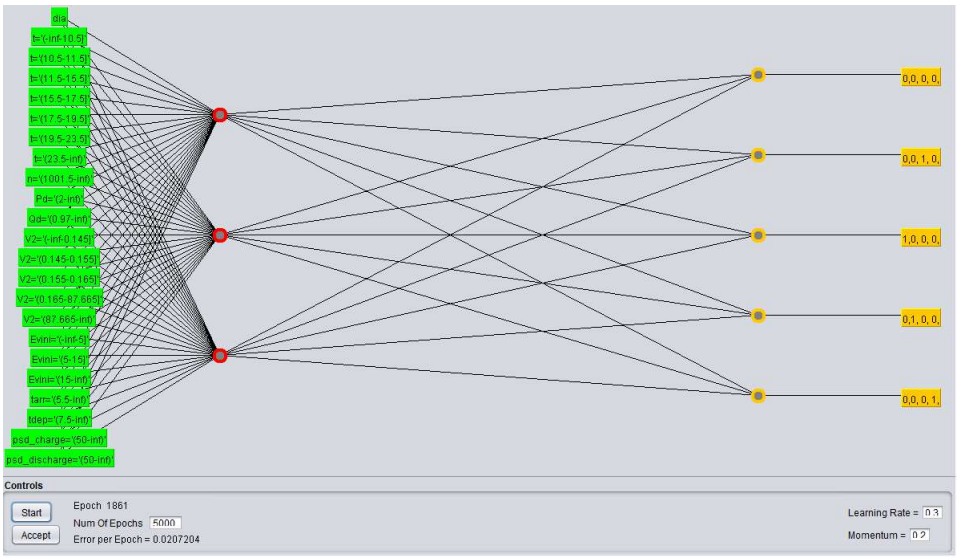
Figura 61 - RNA MLP com 14 camadas ocultas.



Fonte: Próprio Autor

Na Figura 62 é ilustrada uma RNA MLP com 3 camadas ocultas, o número total de iterações para esta RNA é de 5000.

Figura 62 - RNA MLP com 3 camadas ocultas.



Fonte: Próprio Autor

O percentual de classificação durante o processo de criação do modelo de arquitetura para as RNAs realizado nesta pesquisa podem ser vistas na Tabela 24.

Tabela 24 - Percentagem de acerto na classificação durante a construção do modelo para diferentes arquiteturas de RNAs MLP

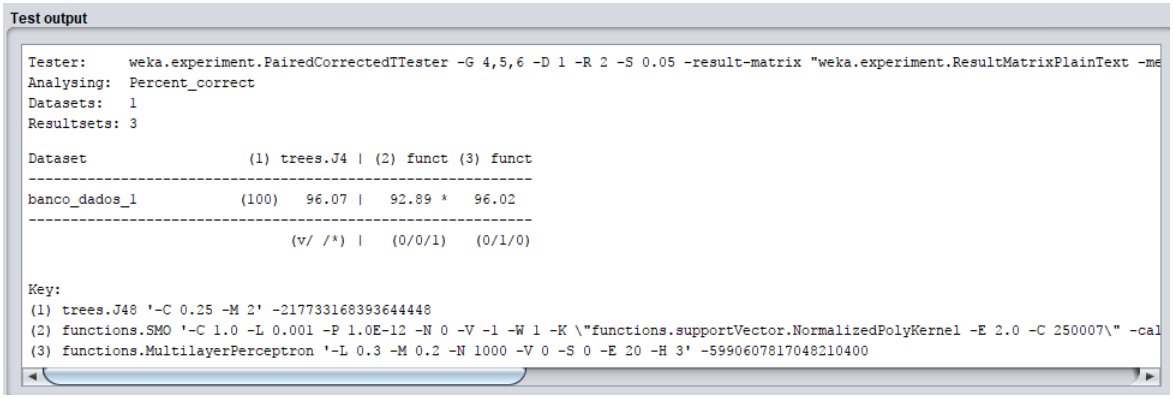
Camadas Ocultas	Iterações	Tempo (segundos)	Classificação (%)
HL	500	8,5	93,02
HL	5000	151,07	96,02
HL	15000	800,93	96,01
3	500	41,45	93,33
3	5000	537,4	93,31
3	15000	996,77	93,37
5	500	49,68	93,83
5	5000	523,63	93,86
5	15000	1422,87	93,87

Fonte: Próprio Autor

Mesmo realizando variação entre o número de camadas ocultas e número de iterações, é notável que o percentual de correta classificação não apresenta muitas mudanças. Na Tabela 24 pode ser visto que a quantidade de iterações e de camadas ocultas tem uma relação direta com o aumento do tempo de criação de cada modelo de RNA.

Na Figura 63 é ilustrada a comparação de percentual de classificação obtido pelos algoritmos de RNAs, ADs e MVS utilizados no desenvolvimento desta pesquisa. Nesta figura pode-se observar que o algoritmo que apresentou melhor resultado foi o algoritmo de ADs com um percentual de correta classificação de 96,07%. Em segundo lugar estão as RNA que apresentaram o percentual de correta classificação de 96,02% e em último lugar estão as MVS que apresentaram um percentual de correta classificação de 92.89%.

Figura 63 - Percentagem de classificação para os diferentes algoritmos que compõem os SIs

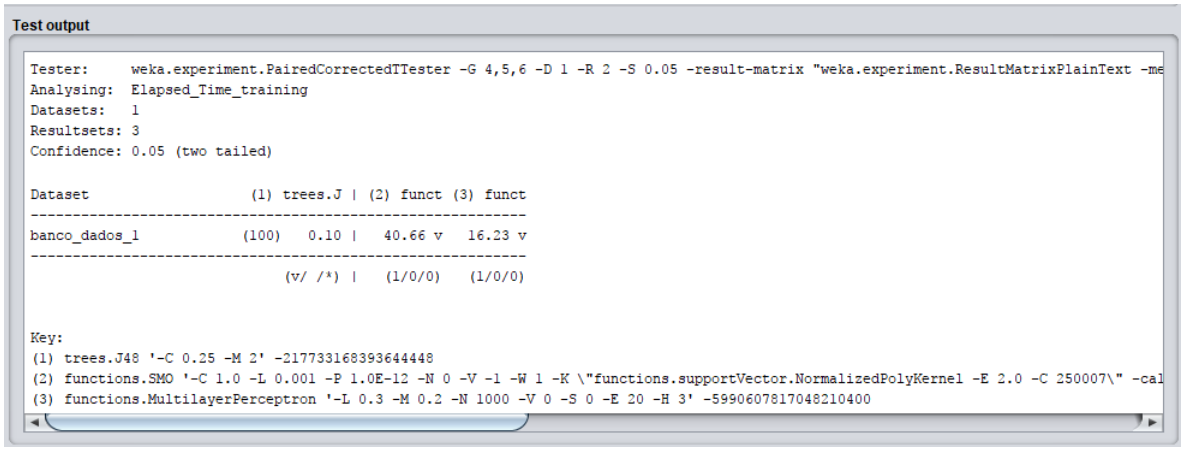


Fonte: Próprio Autor

Na Figura 64 é ilustrada a comparação de tempo de treinamento gasto pelos algoritmos de RNAs, ADs e MVS utilizados no desenvolvimento desta pesquisa. Nesta figura pode-se observar que o algoritmo que apresentou menor tempo de treinamento foi o algoritmo de ADs

com um tempo de 0,10 segundos. Em segundo lugar estão as RNAs que precisaram de 16,23 segundos para realizar o treinamento e em ultimo lugar estão as MVS precisaram para seu treinamento de 40,66 segundos. Portanto pode-se afirmar que o algoritmo que realiza o treinamento de forma mais rápida é o algoritmo de ADs.

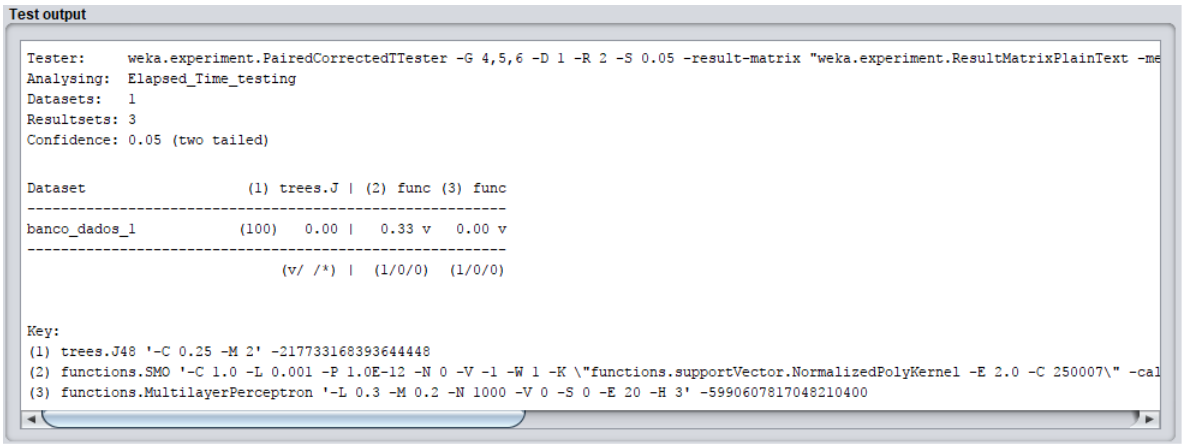
Figura 64 - Comparação de tempo de treinamento entre os algoritmos que compõem os SIs



Fonte: Próprio Autor

Na Figura 65 é ilustrada a comparação de tempo de validação gasto pelos algoritmos de RNAs, ADs e MVS utilizados no desenvolvimento desta pesquisa. Nesta figura pode se observar que o algoritmo que apresentou menor tempo de validação foi o algoritmo de ADs com um tempo de 0,001 segundos. Em segundo lugar estão as RNAs que precisaram de 0,003 segundos para realizar a etapa de validação e em ultimo lugar estão as MVS precisaram para sua validação de 0,33 segundos. Portanto pode-se afirmar que o algoritmo que realiza a validação de forma mais rápida são as ADs e as RNAs, ficando em ultimo lugar as MVS.

Figura 65 - Comparação de tempo de validação entre os algoritmos que compõem os SIs



Fonte: Próprio Autor

Segundo os tempos de validação e o percentual de acertos que apresentam cada um dos algoritmo que compõem os SIs, pode-se afirmar que estas metodologias permitem automatizar o CCIVEs e de DA mediante o uso de SI.

6.4 CONCLUSÕES DO CAPÍTULO

A comparação entre os resultados obtidos na solução de um modelo MPLIM tendo como objetivo coordenação de carga de VEs e DA no SDEE, e os resultados obtidos usando os SIs levam a considerar o uso destas metodologias como alternativas autónomas, rápidas e eficientes.

Os resultados obtidos pelos algoritmos que compõem os SI apresentam erros baixos em comparação com os resultados fornecidos por um MPLIM, o que mostra a utilidade do método proposto.

De acordo com os resultados apresentados pelos métodos baseados em SIs, pode-se afirmar que, para as condições de teste consideradas, as ADs apresentaram um melhor desempenho com maior percentual de acerto na previsão de CCIVEs.

Mesmo que as menores percentagens de acerto, foram apresentadas pelos algoritmos de RNAs e de MVSs, estas percentagens podem ser consideradas como aceitáveis dado que a diferença é consideravelmente pequena.

Para os três algoritmos que compõem os SIs, realizaram-se testes para medir o tempo na etapa de criação do modelo de cada algoritmo, tempo de duração da etapa de teste e da etapa de validação. Segundo os resultados obtidos, a etapa que precisa de maior tempo para sua execução é a etapa de criação do modelo. Uma vez que os modelos estão criados, as etapas de treinamento e de validação precisam de pouco tempo, portanto esta metodologia pode ser indicada para tratar o problema de CCIVE e dos DA nos SDEE em tempo real.

Os tempos de validação dos três algoritmos que compõem os SIs, foram de no máximo 0,33 segundos, portanto podem ser definidos como aceitáveis quando comparado com os tempos de execução necessários de um modelo matemático.

O processo de tomada de decisão pode ser realizado sem precisar da representação detalhada do modelo físico da rede, pois, dados de medições podem ser utilizados como entrada para o treinamento do algoritmos que compõem os SIs.

7 CONCLUSÕES E TRABALHOS FUTUROS

Nesta tese foi utilizada a IC mediante o uso dos algoritmos de AM para solucionar problemas de grande complexidade na área dos Sistemas Elétricos de Potência, como é o caso do Controle centralizado Volt-VAR em modernos Sistemas de Distribuição de Energia Elétrica; a detecção de fraudes nas redes de Distribuição de Energia Elétrica; a localização de faltas em Linhas de Transmissão de Energia Elétrica; e a Coordenação de carga inteligente de veículos elétricos e dispositivos de armazenamento considerando a tecnologia V2G. Dentre os algoritmos de AM utilizados como alternativa de solução encontram-se as RNAs, as MVS e as ADs. Para realizar o gerenciamento adequado dos algoritmos foi usada a plataforma de AM WEKA.

Os resultados obtidos durante as simulações mostram que o software WEKA é uma poderosa ferramenta, que pode ser usada para gerenciar eficientemente os algoritmos de AM, utilizados para resolver problemas de grande complexidade matemática na área dos Sistemas Elétricos de Potência. Estes problemas se caracterizam por possuir uma grande base de dados e um número elevado de variáveis, o que dificulta a tomada de decisões e a obtenção de resultados em tempos computacionais razoáveis quando solucionados através de modelos matemáticos. Portanto, a metodologia proposta surge como uma alternativa interessante que permite obter resultados satisfatórios em menor tempo.

Pode-se concluir que os principais aspectos a serem considerados na hora de resolver um problema usando algoritmos de AM através da plataforma WEKA, são os seguintes:

- Criação da base de dados a partir de um conjunto de medições, ou resultados de um modelo matemático.
- Definição do conjunto de variáveis de interesse do problema (entrada e saída).
- Obtenção de uma porcentagem de aprendizagem durante a etapa de treinamento superior a 95%. Cabe salientar que esta etapa é de maior complexidade, devido aos tempos de simulação e o esforço computacional necessário para construir o modelo de aprendizado.
- Validar os resultados do modelo de aprendizado obtido pelo WEKA com os resultados usados como referência.

De acordo com os resultados obtidos no desenvolvimento da tese, pode-se concluir que o nível de aprendizado de cada algoritmo depende do problema que está-se tratando. Para o problema de Controle centralizado Volt/VAr o algoritmo que teve maior percentagem de aprendizado foi a MVS, apresentando 98,45% de classificação correta. Para o problema de detecçaõ

de fraudes o algoritmo de ADs apresentaram um percentual de aprendizado de 100%. No problema de localização de faltas, o algoritmo de ADs apresentou 96,54% no seu percentual de aprendizado. Por último, no problema de CCIVEs e de DA, o algoritmo com maior nível de aprendizado foi o de AD, seguido das MVS com um percentual de aprendizado de 96,07% e 96,02%, respetivamente.

Considerando os tempos de construção dos modelos dos algoritmos de AM pode-se concluir que, mesmo que os algoritmos que compõem os SIs precisaram de um período de tempo razoável, na hora de se realizar o processo de validação dos mesmos, os tempos de resposta são de no máximo 1 segundo, o que leva a pensar que estas metodologias podem ser catalogadas como aceitáveis quando comparadas com métodos matemáticos utilizados tanto para o controle de magnitude de tensão e de potência reativa (Volt-VAr) em tempo real, quanto na coordenação de carga de VE.

Por outro lado, pode ser concluído que, realizar o processo de localização de faltas nos STEE utilizando SI garante obter resultados mais exatos e em menor tempo, o que leva à diminuição dos tempos de localização e correção da falta, garantindo um melhor serviço por parte das empresas de transmissão de energia elétrica aos seus clientes. Isto quer dizer, que podem ser implementadas dentro do contexto das Redes Inteligentes para resolver o problema de faltas em tempo real.

Segundo os resultados obtidos no desenvolvimento da tese, conclui-se que, pode ser considerado o uso destas metodologias como alternativas rápidas e eficientes de solução de problemas da área dos Sistemas Elétricos de Potência, dado que a comparação entre os resultados obtidos na solução de um modelo MPLIM tendo como objetivo a minimização das perdas do sistema de distribuição, e os resultados obtidos usando os SIs, mostram erros de previsão baixos. Além disso, considerando que poucos trabalhos têm sido desenvolvidos nesta área, as metodologias propostas surgem como ferramenta adicional a ser considerada nas futuras pesquisas que abrangem a modernização e o controle dos SDEE e dos STEE, como é o caso das Redes Inteligentes.

Por outro lado, é possível concluir que, considerando que os algoritmos de IC são tolerantes a ruídos, é possível ter dados de entrada com perturbações e mesmo assim obter um resultado final aceitável, o que leva a pensar que a utilização de métodos de IC são vantajosos dado que permitem realizar o aprendizado das atividades normalmente realizadas por controladores do sistema (humanos), sem que fatores externos afetem a tomada de decisões. Esta afirmação pode ser validada a partir dos resultados obtidos, os quais apresentam erros baixos em comparação com os resultados fornecidos por um modelo matemático, o que mostra a utilidade dos métodos proposto.

Um outro fator a destacar é que esta tese permite concluir que as metodologias desenvolvidas são um importante aporte na área dos Sistemas Elétricos de Potência, pois, foram resolvidos diversos problemas de forma eficiente, mesmo existindo aspetos e/ou fatores limitantes presen-

tes em outras abordagens. Embora seja necessário ter o conhecimento de um especialista na área de determinado problema para sua modelagem, os SI depois de treinados, apresentam resultados aceitáveis em tempos de resposta, quando comparados com o tempo de reação que pode precisar o especialista humano para tomar uma decisão baseado na sua experiência.

Considerando a capacidade de adaptação dos algoritmos de AM pode-se concluir que, os algoritmos de AM podem ser treinados quantas vezes for necessário caso a topologia da rede mude, caso exista um aumento na quantidade de clientes de uma empresa, caso aumentem as variáveis de um determinado problema, caso algum acontecimento novo surgir, mesmo assim, o treinamento necessário para um determinado algoritmo de AM não tem como ser comparado com o tempo de aprendizado de um especialista humano.

De modo geral, pode-se concluir que a utilização de IC permite representar de forma adequada problemas reais sem limitações causadas pela falta ou excesso de medições. Além disso, o processo de tomada de decisões pode ser realizado sem precisar da representação detalhada do modelo físico da rede, pois, dados de medições podem ser utilizados como entrada para o treinamento do SI.

Durante o desenvolvimento de cada capítulo desta tese, foram feitas observações relevantes que devem ser consideradas em trabalhos futuros.

No Capítulo 3:

- Testar outros algoritmos de aprendizado de máquina para determinar qual é mais eficiente na realização do controle centralizado Volt/VAr em tempo real.
- Utilizar aprendizado por reforço e Deep Learning para realizar controle centralizado Volt/VAr em tempo real.

No Capítulo 4:

- Testar a metodologia proposta em dados reais de uma empresa do setor elétrico brasileiro, considerando uma nova variável, como é o caso da variação climática.
- Aplicar Big-Data utilizando informações de medidores inteligentes em tempo real na detecção de clientes suspeitos de cometer algum tipo de fraude.
- Testar outros algoritmos de aprendizado de máquina para determinar qual é mais eficiente na realização do processo de mineração de dados para detecção de clientes suspeitos de cometer algum tipo de fraude.

No Capítulo 5:

- Testar a metodologia proposta em dados reais de uma empresa do setor elétrico brasileiro, considerando a localização de faltas em longas linhas de transmissão.
- Testar outros algoritmos de aprendizado de máquina para determinar qual é mais eficiente na realização da localização da falta em tempo real.

No Capítulo 6:

- Utilizar além das medições do sistema, variáveis estocásticas que ajudem a definir um perfil de uso dos VEs por parte dos donos, e assim obter um SI mais preciso na hora de realizar a CCIVEs e dos DA.
- Utilizar Big-Data para realizar o CCIVEs em um SDEE considerando medições obtidas em tempo real.
- Treinar um SI para que, mediante o uso de um aplicativo móvel, o dono do VE consiga ver em tempo real, uma distribuição ótima das estações de carregamento, evitando um alto índice de coincidência em cada um destes lugares, e diminuindo o tempo de espera por parte dos donos dos VEs.
- Treinar um SI para que permita realizar a predição dos veículos que podem chegar nas próximas horas para serem carregados.

REFERÊNCIAS

AGÊNCIA NACIONAL DE ENERGIA ELÉTRICA - ANEEL. *Condições gerais de fornecimento de energia elétrica, resolução 456*. Brasília, DF, ago. 2000. 57 p.

AGÊNCIA NACIONAL DE ENERGIA ELÉTRICA - ANEEL. *Metodologia de tratamento regulatório para perdas não técnicas de energia elétrica*. Brasília, DF, ago. 2015. 60 p. Nota técnica 106.

ALVES, A. P.; LIMA, A.; SOUZA, S. M. Fault location on transmission lines using complex-domain neural networks. *International Journal of Electrical Power and Energy Systems*, London, v. 43, n. 1, p. 720–727, Fourth 2012. ISSN 0142-0615.

ARAÚJO, A. C. *Considerações sobre as perdas na distribuição de energia eléctrica no Brasil*. 125 f. Tese (Doutorado em Engenharia Elétrica) — Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2006.

ARAÚJO, R.; MEIRA, P.; ALMEIDA, M. Algorithms for operation planning of electric distribution networks. *Journal of Control, Automation and Electrical Systems*, Heidelberg, v. 54, n. 3, p. 154 – 162, Jul. 2013.

ASSOCIAÇÃO BRASILEIRA DE DISTRIBUIÇÃO DE ENERGIA - ABRADE. *CODIGO 08-05 perdas comerciais*. Brasília, DF, ago. 2017.

BELLMAN, R. *An introduction to artificial intelligence: can computers think?* [S.l.]: Boyd & Fraser, 1978. ISBN 9780878350667.

BILGIN, B.; MAGNE, P.; MALYSZ, P.; YANG, Y.; PANTELIC, V.; PREINDL, M.; KOROBKINE, A.; JIANG, W.; LAWFORD, M.; EMADI, A. Making the case for electrified transportation. *IEEE Transactions on Transportation Electrification*, Piscataway, v. 1, n. 1, p. 4–17, June 2015.

BISHOP, C. M. *Pattern recognition and machine learning*. [S.l.]: Springer-Verlag, 2007. 738 p.

BRADY, P.; DAI, C.; BAGHZOUZ, Y. Need to revise switched capacitor controls on feeders with distributed generation. *IEEE PES T&D Conf. Expo*, Dallas, v. 2, p. 590 – 594, Apr. 2003.

BRAHMA, S. M.; GIRGIS, A. A. Fault location on a transmission line using synchronized voltage measurements. *IEEE Transactions on Power Delivery*, Piscataway, v. 19, n. 4, p. 1619–1622, Oct 2004. ISSN 0885-8977.

BUSH, S. F. *Machine intelligence in the grid*. [S.l.]: Wiley-IEEE, 2013. 576 p. ISBN 9781118820216.

CABRAL, J. E.; PINTO, J. O.; GONTIJ, E. M.; REIS, J. Rough sets based fraud detection in electrical energy consumers rough sets based fraud detection in electrical energy consumers. *WSEAS International Conference on Mathematics and Computers in Phisics*, Cancun, Mexico, n. 1–4, Abr. 2004.

CAO, Y.; MIAO, Y.; MIN, G.; WANG, T.; ZHAO, Z.; SONG, H. Vehicular-publish/subscribe (v-p/s) communication nabled on-the-move ev charging management. *IEEE Communications Magazine*, Piscataway, v. 54, n. 12, p. 84–92, December 2016. ISSN 0163-6804.

CAO, Y.; WANG, T.; KAIWARTYA, O.; MIN, G.; AHMAD, N.; ABDULLAH, A. H. An ev charging management system concerning drivers trip duration and mobility uncertainty. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, Piscataway, v. PP, n. 99, p. 1–12, Nov. 2017. ISSN 2168-2216.

CHEN-CHING, L.; PIERCE, D. A.; SONG, H. Intelligent system applications to power systems. *IEEE Computer Applications in Power*, Piscataway, v. 10, n. 4, p. 21–22, 24, Oct 1997. ISSN 0895-0156.

COMPANHIA PARANAENSE DE ENERGIA-COPEL. *Relatório de sustentabilidade COPEL 2016*. Curitiba, ago. 2016. 99 p.

CPLEX DIVISION. *CPLEX optimization subroutine library guide and reference*. New York, ago. 2009. 952 p.

DEILAMI, S.; MASOUM, A. S.; MOSES, P. S.; MASOUM, M. A. S. Real-time coordination of plug-in electric vehicle charging in smart grids to minimize power losses and improve voltage profile. *IEEE Transactions on Smart Grid*, Piscataway, v. 2, n. 3, p. 456–467, Sept 2011. ISSN 1949-3053.

DELGADO, J. *Sistema de informação de apoio a detecção de perdas de energia eléctrica - O caso da Electra*. 140 f. Tese (Doutorado em Engenharia Eletrônica) — Universidade de Aveiro, Santiago, 2010.

DIAGRAMA unifilar da rede de transmissão do sistema Colombiano. Medellin: [s.n.], 2016. 98 p. Dados obtidos de uma empresa cujo nome é confidencial.

ELLER, N. A. *Arquitectura de informação para gestão de perdas comerciais de energia elétrica*. 114 f. Tese (Doutorado em Engenharia de Produção) — Universidade Federal de Santa Catarina, Santa Catarina, 2003.

ENGELS, R.; THEUSINGER, C. Using a data metric for offering preprocessing advice in data mining applications,. In: EUROPEAN CONFERENCE ON ARTIFICIAL INTELLIGENCE ECAI,. Brighton, 1998. *Proceedings*... Brighton: ECAI, 1998, p. 430–434.

FANG, X.; MISRA, S.; XUE, G.; YANG, D. Smart grid. the new and improved power grid: A survey. *IEEE Communications Surveys Tutorials*, Piscataway, v. 14, n. 4, p. 944–980, Fourth 2012. ISSN 1553-877X.

FAYBISOVICH, V.; KHOROSHEV, M. I. Frequency domain double-ended method of fault location for transmission lines. In: TRANSMISSION AND DISTRIBUTION CONFERENCE AND EXPOSITION. IEEE/PES. Chicago, 2008. *Proceedings*... Piscataway: IEEE, 2008, p. 1–6. ISSN 2160-8555.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. *From data mining to knowledge discovery: an overview, in advances in knowledge discovery in databases*. Cambridge: MIT, 1996.

FENG, D.; XIANGJUN, Z.; CHAO, Y.; XIAO'AN, Q.; ZHIHUA, W. Novel traveling wave location algorithm for transmission network based on information fusion technology. In: INTERNATIONAL CONFERENCE ON INTELLIGENT COMPUTATION TECHNOLOGY AND AUTOMATION (ICICTA). [S.l.], 2008. *Proceedings*... Piscataway: IEEE, 2008, p. 1091–1095.

FILHO, J. R.; GONTIJO, E. M.; DELAIBA; MAZINA, E.; CABRAL, J. E.; PINTO, J. O. Fraud identification in electricity company customers using decision tree. In: IEEE INTERNATIONAL CONFERENCE ON SYSTEMS, MAN AND CYBERNETICS (IEEE CAT. NO.04CH37583). [S.l.], 2004. *Proceedings*... Piscataway: IEEE, 2004, p. 3730–3734. ISSN 1062-922X.

FILHO, M. *Sistema inteligente para tomada rápida de decisões nos sistemas elétricos*. 1110 f. Tese (Doutorado em Engenharia Elétrica) — Universidade Federal de Itajubá, Itajubá, 2006.

FORD, F. N. Decision support systems and expert systems: a comparison. *Information & Management*, Elsevier, v. 8, n. 1, p. 21–26, 1985.

FOURER, R.; GAY, D. M.; KERNIGHAN, B. W. *AMPL: A modeling language for mathematical programming*. 2. ed. [S.l.]: Pacific Grove, 2002. 540 p.

FULCZYK, M.; BALCEREK, P.; IZYKOWSKI, J.; ROSOLOWSKI, E. Two-end unsynchronized fault location algorithm for double-circuit series compensated lines. In: POWER AND ENERGY SOCIETY GENERAL MEETING - CONVERSION AND DELIVERY OF ELECTRICAL ENERGY IN THE 21 CENTURY. Pittsburgh, PA, 2008. *Proceedings*... Piscataway: IEEE, 2008, p. 1–9. ISSN 1932-5517.

GAMA, J.; FACELI, K.; LORENA, A. C.; CARVALHO, A. C. *Inteligência artificial: uma abordagem de aprendizado de máquina*. Rio de Janeiro: Grupo Gen - LTC, 2011. ISBN 9788521618805. Disponível em: <https://books.google.com.br/books?id=4DwelAEACAAJ>.

GHARAVI, H.; GHAFURIAN, R. Smart grid: the electric energy system of the future. *Proceedings of the IEEE*, Piscataway, v. 99, n. 6, p. 917–921, Jun. 2011.

GOLDSCHMIDT, R.; PASSOS, E. *Data mining: um guia prático*. Rio de Janeiro: Elsevier, 2005. 265 p.

GOLDSCHMIDT, R.; PASSOS, E.; BEZERRA, E. *Data mining: conceitos, técnicas, algoritmos, orientações e aplicações*. 2. ed. Rio de Janeiro: Elsevier, 2015. 296 p. ISBN 978-85-352-7822-4.

GOMES, L.; FARIA, P.; MORAIS, H.; VALE, Z.; RAMOS, C. Distributed, agent-based intelligent system for demand response program simulation in smart grids. *IEEE Intelligent Systems*, Piscataway, v. 29, n. 1, p. 56–65, Jan 2014. ISSN 1541-1672.

GONÇALVES, R.; ALVES, R. P.; FRANCO, J.; RIDER, M. J. Operation planning of electrical distribution systems using a mixed integer linear model. *Journal of Control, Automation and Electrical Systems*, Springer, v. 24, n. 5, p. 668–679, 2013.

GONÇALVES, R.; ALVES, R. P.; RIDER, M. J. Um modelo linear inteiro misto para o planejamento da operação de redes de distribuição de energia elétrica. In: CONGRESSO BRASILEIRO DE AUTOMÁTICA-CBA, 2012,. Goiania, 2012. *Anais*... Goiania: CBA, 2008.

GOPALAKRISHNAN, A.; KEZUNOVIC, M.; MCKENNA, S. M.; HAMAI, D. M. Fault location using the distributed parameter transmission line model. *IEEE Transactions on Power Delivery*, Piscataway, v. 15, n. 4, p. 1169–1174, Oct 2000. ISSN 0885-8977.

HAN, J.; MICHELINE, K.; JIAN, P. *Data mining: concepts and techniques*. 3. ed. Boston: Morgan Kaufmann, 2012. 703 p. (The Morgan Kaufmann Series in Data Management Systems). ISBN 978-0-12-381479-1.

HAYKIN, S. *Neural networks: a comprehensive foundation*. 3. ed. Ontario Canada: Prentice Hall, 2008. 842 p.

HEBB, D. O. *The organization of behavior: a neuropsychological theory*. 2. ed. New York: Psychology Press, 2008. 335 p.

HERNÁNDEZ-ORALLO, J.; DOWE, D. L. Measuring universal intelligence: towards an anytime intelligence test. *Artificial Intelligence*, Valencia, v. 174, n. 18, p. 1508 – 1539, 2010. ISSN 0004-3702.

HU, X.; MURGOVSKI, N.; JOHANNESSON, L. M.; EGARDT, B. Optimal dimensioning and power management of a fuel cell battery hybrid bus via convex programming. *IEEE-ASME Transactions on Mechatronics*, Piscataway, v. 20, n. 1, p. 457–468, Feb 2015. ISSN 1083-4435.

INTERNATIONAL ENERGY AGENCY - IEA. *World energy statistics 2017*. França, Nov. 2017. 22 p. Disponível em: <https://www.iea.org/statistics/>. Acesso em: 13 Set. 2017.

JAFARI, M.; GAUCHIA, A.; ZHANG, K.; GAUCHIA, L. Simulation and analysis of the effect of real-world driving styles in an ev battery performance and aging. *IEEE Transactions on Transportation Electrification*, Piscataway, v. 1, n. 4, p. 391–401, Dec 2015.

JAHANGIRI, P.; ALIPRANTIS, D. C. Distributed volt/var control by pv inverters. *IEEE Transaction on Power Systems*, New York, v. 28, n. 3, p. 3429–3439, Aug. 2013.

KARFOPOULOS, E. L.; HATZIARGYRIOU, N. D. A multi-agent system for controlled charging of a large population of electric vehicles. *IEEE Transactions on Power Systems*, Piscataway, v. 28, n. 2, p. 1196–1204, May 2013. ISSN 0885-8950.

KAWADY, T.; STENZEL, J. Investigation of practical problems for digital fault location algorithms based on emtp simulation. *IEEE/PES. Transmission and Distribution Conference and Exhibition 2002*, Yokohama, v. 1, p. 118 – 123, Oct. 2002.

KERSTING, W. H. *Distribution system modeling and analysis, regulation of voltages*. 2. ed. Las Cruzes, Nuevo Mexico: CRC, 2007. 425 p.

KIM, I. S. Nonlinear state of charge estimator for hybrid electric vehicle battery. *IEEE Transactions on Power Electronics*, v. 23, n. 4, p. 2027–2034, July 2008. ISSN 0885-8993.

KIM, I. S. A technique for estimating the state of health of lithium batteries through a dual-sliding-mode observer. *IEEE Transactions on Power Electronics*, Piscataway, v. 25, n. 4, p. 1013–1022, April 2010. ISSN 0885-8993.

KURZWEIL, R. *The age of intelligent machines*. [S.l.]: MIT, 1990. 565 p.

LABERGE, R. *The data warehouse mentor: practical data warehouse and business intelligence insights*. New York: McGraw-Hill, 2011. 416 p.

LAMBERT-TORRES, G.; RIBEIRO, G.; COSTA, C.; SILVA, A. A. da; QUINTANA, V. Knowledge engineering tool for training power-substation operators. *IEEE Transactions on Power Delivery*, Piscataway, v. 12, n. 2, p. 694–699, Apr 1997. ISSN 0885-8977.

LIANG, R. H.; CHENG, C. K. Dispatch of main transformer ultc and capacitors in a distribution system. *IEEE Transactions on Power Delivery*, Piscataway, v. 16, n. 4, p. 625–630, Oct 2001. ISSN 0885-8977.

LIANG, R. H.; WANG, Y. S. Fuzzy-based reactive power and voltage control in a distribution system. *IEEE Transactions on Power Delivery*, Piscataway, v. 18, n. 2, p. 610–618, April 2003. ISSN 0885-8977.

LIANG, R. H.; WANG, Y. S. A universal fault location technique for n-terminal transmission lines. *IEEE Transactions on Power Delivery*, Piscataway, v. 23, n. 3, p. 1366–1373, July 2008. ISSN 0885-8977.

LIANG, Y.; WANG, G.; LI, H. A novel fault location for transmission line by the combination of particle swarm optimization and least squares method. In: INTERNATIONAL POWER ENGINEERING CONFERENCE (IPEC 2007). Singapore, 2007. *Proceedings*...Piscataway: IEEE, 2007, p. 1151–1155. ISSN 1947-1262.

LIAO, Y.; KEZUNOVIC, M. Optimal estimate of transmission line fault location considering measurement errors. *IEEE Transactions on Power Delivery*, Piscataway, v. 22, n. 3, p. 1335–1341, July 2007. ISSN 0885-8977.

LIN, X.; BOGDAN, P.; CHANG, N.; PEDRAM, M. Machine learning-based energy management in a hybrid electric vehicle to minimize total operating cost. In: INTERNATIONAL CONFERENCE ON COMPUTER-AIDED DESIGN (ICCAD). Austin, 2015. *Proceedings*... Piscataway: IEEE/ACM, 2015, p. 627–634.

LIN, X.; MAO, P.; WENG, H.; WANG, B.; BO, Z. Q.; KLIMEK, A. Study on fault location for high voltage overhead transmission lines based on neural network system. In: INTERNATIONAL CONFERENCE ON INTELLIGENT SYSTEMS APPLICATIONS TO POWER SYSTEMS. Toki Messe, 2011. *Proceedings*...Toki Messe, 2011, p. 1–5.

MAHMOOD, A. M.; SATULURI, N.; KUPPA, M. R. An overview of recent and traditional decision tree classifiers in machine learning. *International Journal of Research and Reviews in Ad Hoc Networks*, New York, v. 1, n. 1, p. 1688–1697, Apr. 2011.

MAJID, I. A.; RAHMAN, R. F.; SETIAWAN, N. A.; CAHYADI, A. I. Electric vehicle battery dynamics modelling using support vector machine. In: JOINT INTERNATIONAL CONFERENCE ON RURAL INFORMATION COMMUNICATION TECHNOLOGY AND ELECTRIC-VEHICLE TECHNOLOGY (RICT ICEV-T). [S.l.], 2013. *Proceedings*...Piscataway: IEEE, 2013, p. 1–3.

MCCULLOCH, M. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, Chicago, v. 5, n. 1, p. 115–133, Apr. 1943.

MEI, K.; ROVNYAK, S. M. Response-based decision trees to trigger oneshot stabilizing control. *IEEE Transactions on Power Systems*, Piscataway, v. 19, n. 1, p. 531–537, May. 2004.

MENG, J.; LUO, G.; GAO, F. Lithium polymer battery state-of-charge estimation based on adaptive unscented kalman filter and support vector machine. *IEEE Transactions on Power Electronics*, Piscataway, v. 31, n. 3, p. 2226–2238, March 2016. ISSN 0885-8993.

METS, K.; D'HULST, R.; DEVELDER, C. Comparison of intelligent charging algorithms for electric vehicles to reduce peak load and demand variability in a distribution grid. *Journal of Communications and Networks*, Seoul, v. 14, n. 6, p. 672–681, Dec 2012. ISSN 1229-2370.

MITCHELL, T. M. *Machine learning*. [S.l.]: McGraw-Hill, 1997. 421 p. (McGraw-Hill International Editions). ISBN 9780071154673.

MOHAPATRA, A.; BIJWE, P. R.; PANIGRAHI, B. K. An efficient hybrid approach for volt/var control in distribution systems. *IEEE Transactions on Power Delivery*, Piscataway, v. 29, n. 4, p. 1780–1788, Aug 2014. ISSN 0885-8977.

MOMOH, J. A. *Electric power distribution, automation, protection, and control*. [S.l.]: CRC, 2017. 378 p.

MULLER, K.; MIKA, S.; RATSCH, G.; TSUDA, K.; SCHOLKOPF, B. An introduction to kernel-based learning algorithms. *IEEE Transactions on Neural Networks*, New York, v. 12, n. 2, p. 181–201, Mar 2001. ISSN 1045-9227.

NIKNAM, T.; ZARE, M.; AGHAEI, J. Scenario-based multiobjective volt/var control in distribution networks including renewable energy sources. *IEEE Transactions on Power Delivery*, New York, v. 27, n. 4, p. 2004–2019, Oct 2012. ISSN 0885-8977.

NILSSON, N. J. *Artificial intelligence: a new synthesis*. New York: Elsevier, 1998. 513 p.

NORTHCOTE-GREEN, J.; WILSON, R. G. *Control and automation of electrical power distribution systems*. New York: CRC, 2006. 488 p.

OBA, C. C.; NUNES, A.; SINKITI, D.; PAPA, J. P. Identification and feature selection of non-technical losses for industrial consumers using the software weka. In: IAS INTERNATIONAL CONFERENCE ON INDUSTRY APPLICATIONS (INDUSCON), 10., 2012,. Fortaleza, 2012. *Proceedings*... Piscataway: IEEE, 2012, p. 1–6.

O'CONNELL, A.; FLYNN, D.; KEANE, A. Rolling multi-period optimization to control electric vehicle charging in distribution networks. *IEEE Transactions on Power Systems*, National Harbor, v. 29, n. 1, p. 340–348, Jan 2014. ISSN 0885-8950.

OLIVEIRA, E.; BRAZ, J.; FERREIRA, D. Um processador de alarme inteligente. *SBA Controle e Automação*, Campinas, v. 4, n. 2, p. 55 – 61, May. 1994.

PARÂMETROS da linha de transmissão. Medellin: [s.n.], 2016. 87 p. Dados obtidos de uma empresa cujo nome é confidencial.

PARK, J. Y.; NAM, S. R.; PARK, J. K. Control of a ultc considering the dispatch schedule of capacitors in a distribution system. *IEEE Transactions on Power Systems*, Piscataway, v. 22, n. 2, p. 755–761, May 2007. ISSN 0885-8950.

PLATT, J. C. Sequential minimal optimization: a fast algorithm for training support vector machines. 1998.

QUEIROGA, R. M. *Uso de técnicas de data mining para detecção de fraudes em energia elétrica*. 147 f. Tese (Doutorado em Informática)) — Universidade Federal do Espírito Santo, Vitória:, 2005.

QUINLAN, J. R. Improved use of continuous attributes in c4.5. *Journal of Artificial Intelligence Research*, New York, v. 4, n. 1, p. 77 – 90, Nov. 1996.

RAILEANU, L.; STOFFEL, K. Theoretical comparison between the gini index and information gain criteria. *Annals of mathematics and artificial intelligence*, p. 77 – 93, Oct. 2004.

REZENDE, S. O.; PUGLIESE, J. B.; AO, F. M. V. *Sistemas baseados em conhecimentos, sistemas inteligentes*. [S.l.]: Editora Manole, 2003.

RICH, E.; KNIGHT, K. *Artificial intelligence*. 2. ed. United States of American: McGraw-Hill, 1991. 640 p.

ROKACH, L.; MAIMON, O. *Data mining with decision trees: theory and applications. Series in machine perception and artificial intelligence*. 2. ed. Singapore: World Scientific Publishing Company, 2008. 244 p.

ROSENBLATT, F. *Principles of neurodynamics: perceptrons and the theory of brain mechanisms*. Washington: Spartan Books, 1962. 616 p.

ROYTELMAN, I.; GANESAN, V. Coordinated local and centralized control in distribution management systems. *IEEE Transactions on Power Delivery*, Piscataway, v. 15, n. 2, p. 718 – 724, Apr. 2000.

RUSSELL, S. J.; NORVIG, P. *Artificial intelligence: a modern approach*. 3. ed. New Jersey: Pearson, 2009. 1152 p.

SABILLÓN, C.; FRANCO, J. F.; RIDER, M. J.; ROMERO, R. A new methodology for the optimal charging coordination of electric vehicles considering vehicle-to-grid technology. *IEEE Transactions on Sustainable Energy*, Piscataway, v. 7, n. 2, p. 596–607, April 2016. ISSN 1949-3029.

SABILLÓN, C. F.; FRANCO, J. F.; RIDER, M. J.; ROMERO, R. A milp model for optimal charging coordination of storage devices and electric vehicles considering v2g technology. In: INTERNATIONAL CONFERENCE ON ENVIRONMENT AND ELECTRICAL ENGINEERING (EEEIC), 15., 2015. Rome, Italy, 2015. *Proceedings*... Piscataway: IEEE, 2015, p. 60–65.

SAHA, M. M.; DAS, R.; VERHO, P.; NOVOSEL, D. Review of fault location techniques for distribution systems. *Power Systems and Communications Infrastructures for the future*, Australia, v. 12, n. 2, p. 2670–2677, 2012.

SHAABAN, M. F.; ISMAIL, M.; EL-SAADANY, E. F.; ZHUANG, W. Real-time pev charging/discharging coordination in smart distribution systems. *IEEE Transactions on Smart Grid*, Piscataway, v. 5, n. 4, p. 1797–1807, July 2014. ISSN 1949-3053.

SHAHID, N.; ALEEM, S. A.; NAQVI, I. H.; ZAFFAR, N. Support vector machine based fault detection amp; classification in smart grids. In: GLOBECOM WORKSHOPS, 2012. [S.l.], 2012. *Proceedings*... Piscataway: IEEE, 2012, p. 1526–1531. ISSN 2166-0077.

SHANNON, C. E.; WEAVER, W. *The mathematical theory of communication*. Illinois: University of Illinois, 1971. 144 p.

SHAWE-TAYLOR, J.; CRISTIANINI, N. *Support vector machines and other kernel-based learning methods*. Cambridge: Cambridge University Press, 2000. 204 p.

SILVA, M. P. S. *Mineração de dados - Conceitos, aplicações e experiência com WEKA*. Rio Janeiro: Escola regional de Informática de Rio de Janeiro, 2004.

SILVA, M. P. S.; ROBIN, J. R. *SKDQL - Uma linguagem declarativa de especificação de consultas e processos para descoberta de conhecimento em bancos de dados e sua implementação*. [S.l.]: UFPE, 2003.

SUNIL, S.; VISHWAKARMA, D. N. Intelligent techniques for fault diagnosis in transmission lines 2014: an overview. In: INTERNATIONAL CONFERENCE ON RECENT DEVELOPMENTS IN CONTROL, AUTOMATION AND POWER ENGINEERING (RDCAPE), 2015. Noida, 2015. *Proceedings*... Piscataway: IEEE, 2015, p. 280–285.

TAN, Y.; WANG, J. A support vector machine with a hybrid kernel and minimal vapnik-chervonenkis dimension. *IEEE Transactions on Knowledge and Data Engineering*, Piscataway, v. 16, n. 4, p. 385–395, April 2004. ISSN 1041-4347.

TANG, L.; RIZZONI, G.; ONORI, S. Energy management strategy for hevs including battery life optimization. *IEEE Transactions on Transportation Electrification*, Piscataway, v. 1, n. 3, p. 211–222, Oct 2015.

TANG, Y.; ZHONG, J.; BOLLEN, M. Aggregated optimal charging and vehicle-to-grid control for electric vehicles under large electric vehicle population. *IET Generation, Transmission Distribution*, Piscataway, v. 10, n. 8, p. 2012–2018, 2016. ISSN 1751-8687.

TAO, X.; RENMU, H.; PENG, W.; DONGJIE, X. Applications of data mining technique for power system transient stability prediction. *Electric Utility Deregulation, Restructuring and Power Technologies*, Beijing, v. 1, n. 1, p. 389–392, Apr. 2004.

THOTA, L. S.; BADAWY, A. S.; CHANGALASETTY, S. B.; GHRIBI, W. Classify vehicles: classification or clusterization? In: INTERNATIONAL CONFERENCE ON CIRCUITS, POWER AND COMPUTING TECHNOLOGIES (ICCPCT), 2015]. Nagercoil, 2015. *Proceedings*... Piscataway: IEEE, 2015, p. 1–7.

TRIPPE, A.; MASSIER, T.; HAMACHER, T. Optimized charging of electric vehicles with regard to battery constraints - case study: Singaporean car park. In: ENERGYTECH, 2013. Cleveland, 2013. *Proceedings*... Piscataway: IEEE, 2013, p. 1–6.

TSO, S. K.; LIN, J. K.; HO, H. K.; MAK, C. M.; YUNG, K. M.; HO, Y. K. Data mining for detection of sensitive buses and influential buses in a power system subjected to disturbances. *IEEE Transactions on Power Systems*, Beijing, v. 19, n. 1, p. 563–568, Feb 2004. ISSN 0885-8950.

VALIQUETTE, B.; TORRES, L.; MUKHEDKAR, D. An expert system based diagnosis and advisor tool for teaching power system operation emergency control strategies. *IEEE Transactions on Power Systems*, Piscataway, v. 6, n. 3, p. 1315–1322, Aug 1991. ISSN 0885-8950.

VAPNIK, V. N. *Statistical learning theory*. New York: Wiley-Interscience, 1998. 768 p.

VASSENA, S.; MACK, P.; ROUSEAUX, P.; DRUET, C.; WEHENKEL, L. A probabilistic approach to power system network planning under uncertainties. In: POWER TECH CONFERENCE, 2003. Bologna, 2003. *Proceedings*... Piscataway: IEEE, 2013, p. 6 pp. Vol.2–.

VENKATESAN, R.; BALAMURUGAN, B. A real-time hardware fault detector using an artificial neural network for distance protection. *IEEE Transactions on Power Delivery*, Piscataway, v. 16, n. 1, p. 75 – 82, Jan. 2001.

VIAWAN, F. A.; KARLSSON, D. Voltage and reactive power control in systems with synchronous machine-based distributed generation. *IEEE Transactions on Power Delivery*, Piscataway, v. 23, n. 2, p. 1079 – 1087, Jan. 2008.

VILLACCI, D.; BONTEMPI, G.; VACCARO, A. An adaptive local learning-based methodology for voltage regulation in distribution networks with dispersed generation. *IEEE Transactions On Power Systems*, Piscataway, v. 21, n. 3, p. 1131 – 1140, Aug. 2006.

VILLALOBOS, J. G. ia; ZAMORA, I.; IN, J. I. S. M.; ASENSIO, F. J.; APERRIBAY, V. Plug-in electric vehicles in electric distribution networks: a review of smart charging approaches. *Renewable and Sustainable Energy Reviews*, Elsevier, Kidlington, v. 38, p. 717–731, 2014. ISSN 1364-0321.

WANG, B.; DONG, X.; BO, Z.; KLIMEK, A. Impedance phase faults location algorithm for uhv transmission lines. In: IEEE/PES TRANSMISSION AND DISTRIBUTION CONFERENCE AND EXPOSITION: LATIN AMERICA, 2008. Bogotá, 2008. *Proceedings*... Piscataway: IEEE, 2008, p. 1–4.

WANG, N.; ARAVINTHAN, V.; DING, Y. Feeder-level fault detection and classification with multiple sensors: A smart grid scenario. In: WORKSHOP ON STATISTICAL SIGNAL PROCESSING (SSP), 2014. Gold Coast, 2014. *Proceedings*... Piscataway: IEEE, 2014, p. 37–40. ISSN 2373-0803.

WANG, R.; XIAO, G.; WANG, P. Hybrid centralized-decentralized (hcd) charging control of electric vehicles. *IEEE Transactions on Vehicular Technology*, Piscataway, v. 66, n. 8, p. 6728–6741, Aug 2017. ISSN 0018-9545.

WERBOS, P. J. Computational intelligence for the smart grid-history, challenges, and oportunities. *IEEE Computational Intelligence Magazine*, Piscataway, v. 6, n. 3, p. 14–21, Aug 2011. ISSN 1556-603X.

WINSTON, P. H. *Artificial intelligence*. 3. ed. New York: Pearson, 1992. 737 p.

WITTEN, I. H.; FRANK, E. *Data mining: practical machine learning tools with JAVA implementations*. United States of America: Morgan Kaufmann Publisher, 2000. 558 p.

WITTEN, I. H.; FRANK, E. *Data mining: practical machine learning tools and techniques*. 3. ed. United States of America: Morgan Kaufmann Publisher, 2011. 664 p.

WITTEN, I. H.; FRANK, E.; HALL, M. A.; PAL, C. J. *Data mining: practical machine learning tools and techniques*. 4. ed. [S.l.]: Elsevier Science, 2016. 654 p. (The Morgan Kaufmann Series in Data Management Systems). ISBN 9780128043578.

WOLLENBERG, B. F. Feasibility study for an energy management system intelligent alarm processor. *IEEE Transactions on Power Systems*, Piscataway, v. 1, n. 2, p. 241 – 247, 1986.

YILMAZ, M.; KREIN, P. T. Review of battery charger topologies, charging power levels, and infrastructure for plug-in electric and hybrid vehicles. *IEEE Transactions on Power Electronics*, Piscataway, v. 28, n. 5, p. 2151–2169, May 2013. ISSN 0885-8993.

ZHANG, F.; LIANG, J.; ZHANG, L.; YUN, Z. A new fault location method avoiding wave speed and based on traveling waves for ehv transmission line. In: THIRD INTERNATIONAL CONFERENCE ON ELECTRIC UTILITY DEREGULATION AND RESTRUCTURING AND POWER TECHNOLOGIES, 2008. Nanjing, 2008. *Proceedings*... Nanjing, 2015, p. 1753–1757.

ZHAO, Y.; ZHANG, Y.; SATULURI, N.; KUPPA, M. R. Comparison of decision tree methods for finding active objects. *National Astronomical Observatories, Advances of Space Research*, Bejing, v. 1, n. 1, p. 1 – 10, Jul. 2007.

APÊNDICE A - WEKA

Neste capítulo é apresentada a plataforma de aprendizagem de máquina Waikato Environment for Knoowledge Analysis (WEKA) utilizada nos testes realizados durante o desenvolvimento desta tese. O apêndice está organizado da seguinte forma: na Seção 8.1 é apresentada a introdução ao WEKA. Na Seção 8.2 são apresentadas as diferentes ferramentas de gerenciamento de algoritmos de aprendizado de máquina mais utilizados em IC. Na Seção 2.3 são apresentadas as principais características de cada algoritmo utilizados no desenvolvimento desta tese. Finalmente, na Seção 2.4 é feita uma descrição da utilização da ferramenta WEKA.

A.1 INTRODUÇÃO

Esta plataforma foi escolhida para gerenciar os SIs, dado que é uma ferramenta que, além de oferecer uma grande lista de algoritmos de aprendizagem de maquina, é uma ferramenta *open source*, isto é, ferramenta de código aberto, o que possibilita realizar mudanças no código para alcançar a arquitetura necessária de algum algoritmo que se esteja utilizando.

Este software criado pela Universidade de Waikato na Nova Zelânda, tem uma extensa variedade de algoritmos de aprendizado de máquinas, desenvolvidos na linguajem de programação Java (orientado a objetos), que são úteis para serem aplicados em bancos de dados fazendo uso das interfazes que oferece. Além disso, o WEKA contém as ferramentas necessárias para realizar transformações sobre os dados, tarefas de classificação, regressões, clustering, associação e visualização. O WEKA está desenhado como uma ferramenta orientada à extensibilidade, pelo que adicionar novas funcionalidades é uma tarefa simples. O WEKA é um software de livre distribuição, multiplataforma, e está conformado por uma série de pacotes de código aberto com diferentes técnicas. Estes pacotes podem ser integrados em qualquer projeto de análise de dados, e até podem estender-se com contribuições dos usuários que desenvolvem novos algoritmos (WITTEN; FRANK, 2011).

Embora o WEKA seja uma ferramenta de qualidade, também apresenta uma grande desvantagem dada pela pouca documentação orientada ao usuário que permita melhorar a usabilidade, isto faz como que a ferramenta seja difícil de entender e de usar. Por ser uma ferramenta desenvolvida em JAVA é independente da arquitetura, por tanto, pode ser executada em qualquer plataforma na qual exista uma máquina virtual JAVA disponível.

A.1.1 Escolha de plataforma de aprendizagem de máquina

Dado que existe uma grande oferta de plataformas de aprendizagem de máquina, naturalmente tem que se proceder a uma seleção de modo a tornar viável uma comparação entre as diferentes técnicas que são utilizadas. Os critérios de escolha da ferramenta passam pelo fato da mesma ter que implementar RNAs, MVSs e ADs. Por outro lado, a ferramenta deve ter uma interface simples, passível de ser experimentada por utilizadores não especializados.

Tabela 25 - Ferramentas que gerenciam SIs.

Ferramenta	RNAs			MVSs	ADs
	NN	MLP	RBF		
Clementine		√	√	√	√
Orange	√			√	√
R	√	√	√	√	√
Tiberius	√			√	√
Weka	√	√	√	√	√

Fonte: Próprio Autor

Na Tabela 25, pode-se verificar que não são muitas as ferramentas possíveis de serem utilizadas nas experiências pretendidas e algumas destas possuem, ainda, particularidades que as tornam impróprias para as experiências. No caso da ferramenta Clementine, por ser comercial e representar um alto valor, além de não disponibilizar versão de demonstração, esta ferramenta foi descartada. Uma situação idêntica é o caso da ferramenta Tiberius que, embora disponibilize a versão de demonstração, é demasiado limitada para os propósitos dos testes que tem que ser realizados. Por sua vez, a ferramenta Orange possui uma interface gráfica muito interessante, no entanto, não possui MVSs adequadas. Portanto, restam as ferramentas R e WEKA. Cabe salientar que a ferramenta R tem uma interface via linha de comando, o que não é propriamente adequada à sua utilização por utilizadores não especializados. Contudo, a ferramenta escolhida para realizar a implementação dos SIs é o WEKA.

A ferramenta WEKA gerência os algoritmos de RNAs, MVSs e ADs. Como interface utiliza um Graphic User Interface (GUI), tendo uma licencia não comercial.

Constata-se que das ferramentas não comerciais, o Weka é uma ferramenta cada dia mais utilizada no desenvolvimento de pesquisas na área de IC aplicada a diversas áreas interdisciplinares. Na literatura podem ser encontrados trabalhos como o apresentado por (OBA et al., 2012) no qual é utilizada a ferramenta WEKA para identificar perdas não técnicas e selecionar as características mais relevantes do problema, considerando informações do perfil de consumidores industriais contidas em um banco de dados de uma empresa concessionária de energia elétrica. Neste trabalho o WEKA é utilizado para realizar a comparação das técnicas de classificação e otimização através de algoritmos inteligentes.

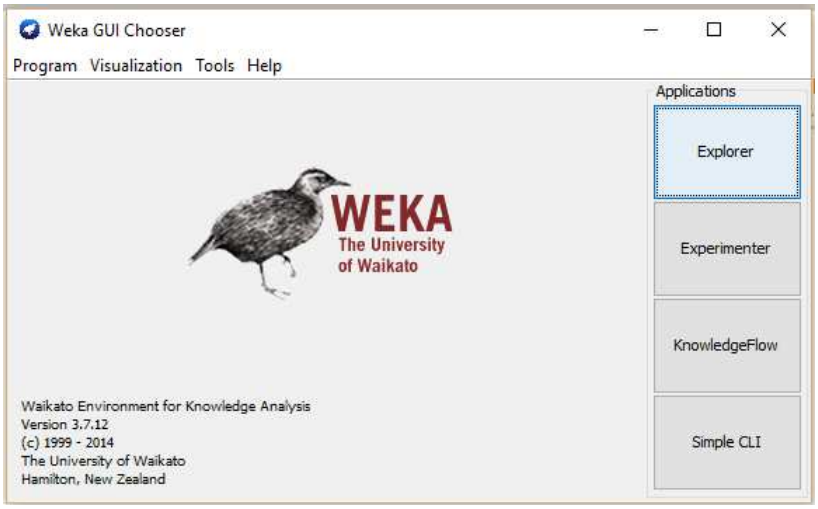
Outras pesquisas utilizam a ferramenta WEKA é apresentada por Thota et al. (2015). Neste trabalho é analisada uma sequência de imagens de cenas de tráfego veicular que são registradas por câmaras inteligentes estacionárias. A mineração de dados é utilizada para automatizar as técnicas de análises de dados e para descobrir relações entre itens como: largura, comprimento, área e perímetro, e posteriormente classificar os veículos com necessidades grandes ou pequenas, para assim, realizar um processo de prevenção de acidentes, diminuir o tráfego, e melhorar a estrutura de vigilância veicular. A ferramenta WEKA foi utilizada para aplicar os algoritmos de Mineração de dados e realizar o processo de classificação veicular.

Essa não é a única aplicação da ferramenta WEKA em áreas interdisciplinares, há uma coleção de artigos publicados das pesquisas desenvolvidas em áreas como: medicina, engenharia da computação, engenharia elétrica, engenharia de controle, engenharia de produção, entre outras. Existem diversas aplicações desenvolvidas fazendo uso da ferramenta computacional WEKA, cada dia seu uso torna-se mais eficiente, isto graças a que é uma ferramenta que esta em constante desenvolvimento, o que faz com que mudanças e melhoras sejam realizadas, possibilitando o uso desta ferramenta para o desenvolvimento de pesquisas cada vez mais complexas.

A.2 DESCRIÇÃO DA FERRAMENTA WEKA

Como já foi mencionado anteriormente, a ferramenta WEKA possui uma interface gráfica que permite o gerenciamento dos algoritmos de IC utilizados durante o desenvolvimento desta pesquisa. A tela inicial que pode ser vista na Figura 66, ilustra ao utilizador uma interface, na qual pode escolher entre quatro possíveis opções, cada uma com suas características específicas.

Figura 66 - Janela inicial do WEKA (GUI Chooser).



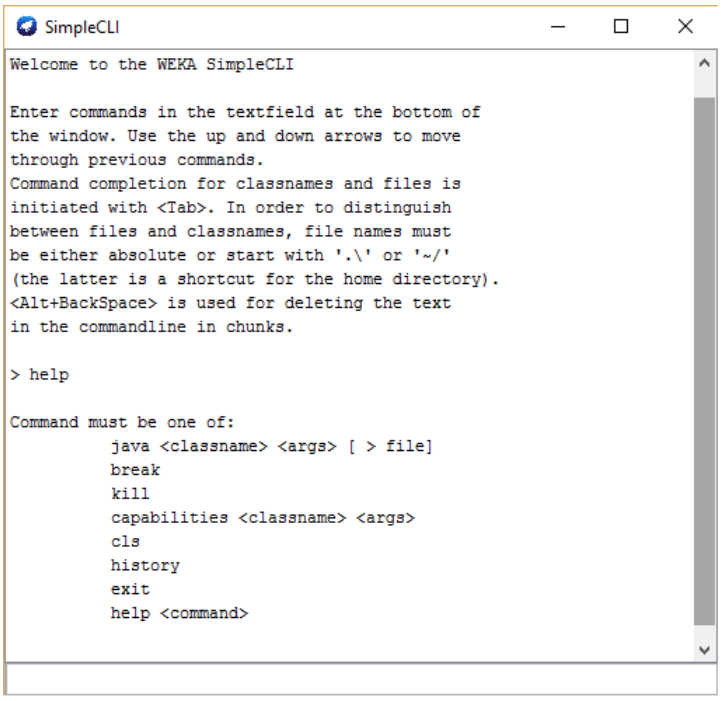
Fonte: Próprio Autor

Cabe salientar que WEKA é um software em desenvolvimento constante e cada uma das suas interfaces evolui separadamente. Portanto, na atualidade cada interface não é completamente equivalente uma com a outra, isto é, existem algumas funcionalidades que só podem ser realizadas utilizando determinada interface.

A.2.1 Simple CLI (Command Line Interface)

Proporciona uma interface de linha de comando onde podem ser executadas linhas de comando do WEKA. Embora, disponibilize todas as funcionalidades, requer um elevado grau de conhecimento dos comandos que podem ser utilizados. Mesmo tendo uma aparência que faz pensar que é uma ferramenta simples, esta interface é muito poderosa e permite realizar qualquer operação suportada pelo WEKA de forma direta; mas seu uso é muito complexo, pois, é necessário um amplo conhecimento da ferramenta WEKA. Sua utilidade é pequena e atualmente é utilizada esta interface como ferramenta de ajuda na fase de testes. A tela do Simple CLI é ilustrada na Figura 67

Figura 67 - Interface de linha de comando do WEKA.



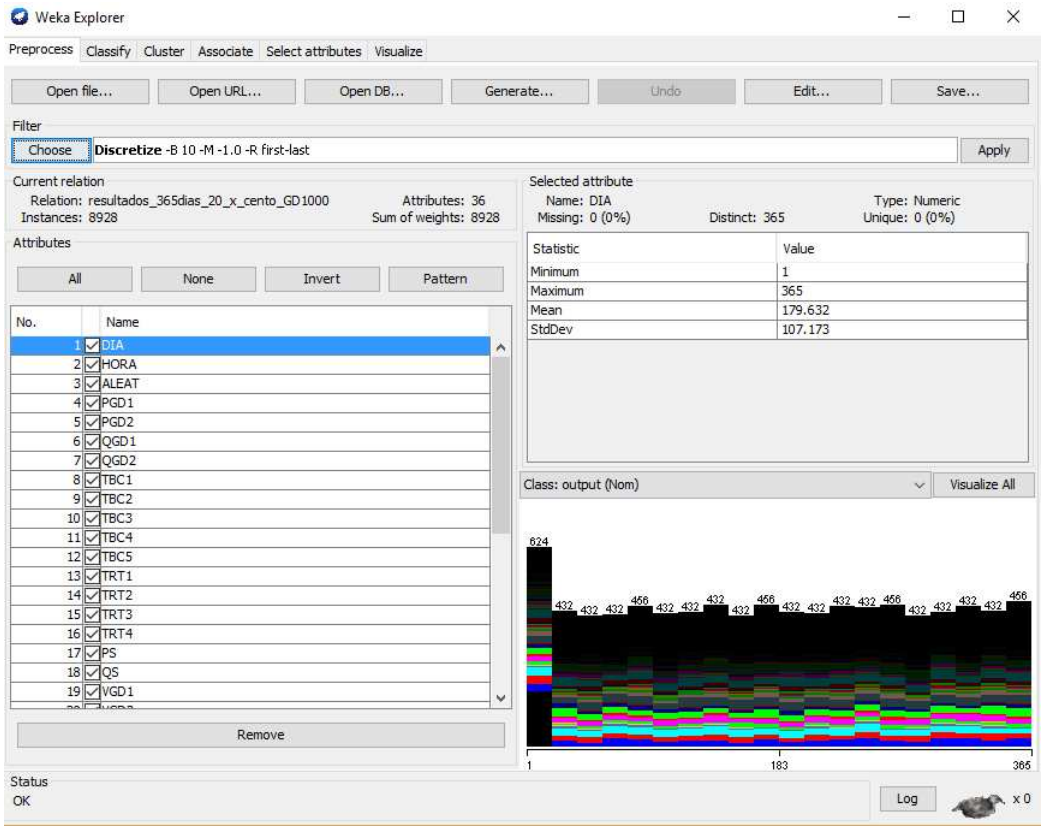
Fonte: Próprio Autor

A.2.2 Explorer

Proporciona um ambiente gráfico de manipulação de dados pela utilização de diferentes algoritmos. É a interface mais fácil de utilizar, guiando ao utilizador através de menus e formulários, impedindo-o e fazer escolhas não aplicáveis, e apresentado *pop-ups* de informação

sobre o preenchimento de cada um dos campos, como pode ser visto na Figura 68 . Além disso, é a interface mais utilizada e a mais descritiva. Esta interface permite realizar operações sobre um único banco de dados. O *Explorer* permite tarefas de: pre-processamento de dados e aplicações de filtros, classificação, clustering, associação, seleção de atributos e visualização de dados estatísticos.

Figura 68 - Ambiente Explorer.



Fonte: Próprio Autor

O *Explorer* oferece a utilização de seis painéis:

- **Preprocess:** Seleção dos dados e preparação (filtro). Inclui as ferramentas e filtros para carregar e manipular os dados.
- **Classify:** Facilidades para aplicar esquemas de classificação e regressão, treinar modelos e avaliar sua presição.
- **Cluster:** Integra vários métodos de agrupamento.
- **Associate:** Inclui umas técnicas de regras de associação.
- **Select Attributes:** Busca supervisionada de subconjuntos de atributos representativos. Permite aplicar várias técnicas para a redução de número de atributos.

- **Visualize:** Neste entorno pode-se estudar o comportamento dos dados fazendo uso das técnicas de visualização.

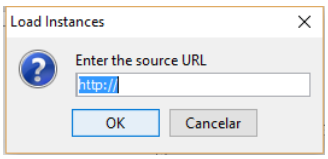
A.2.2.1 Preprocess

O passo inicial para trabalhar com o **Explorer** é definir a origem dos dados. WEKA suporta diferentes bancos de dados que correspondem com os botões que estão em baixo das abas superiores. As diferentes opções são as seguintes:

- **Open File:** Ao dar click sobre este botão aparecerá uma janela de seleção de arquivo. Embora o formato padrão do WEKA seja o ARFF, isto não quer dizer que seja o único que ele aceite , pois, o WEKA internamente tem interpretadores para outros formatos de dados. Estes interpretadores são:
  CSV, arquivos separados por vírgulas ou tabulações, onde a primeira linha contém os atributos;
  C 4.5, arquivo codificado segundo o formato C4.5;
  Instâncias serializadas, WEKA internamente armazena cada amostra dos bancos de dados como uma instância da classe. Esta classe pode ser serializada, pelo que estes objetos podem ser carregados diretamente de um arquivo.

- **Open URL:** com este botão aparece uma janela que permite introduzir um endereço definindo a localização dos dados. Seu menu de opções é ilustrado na Figura 69

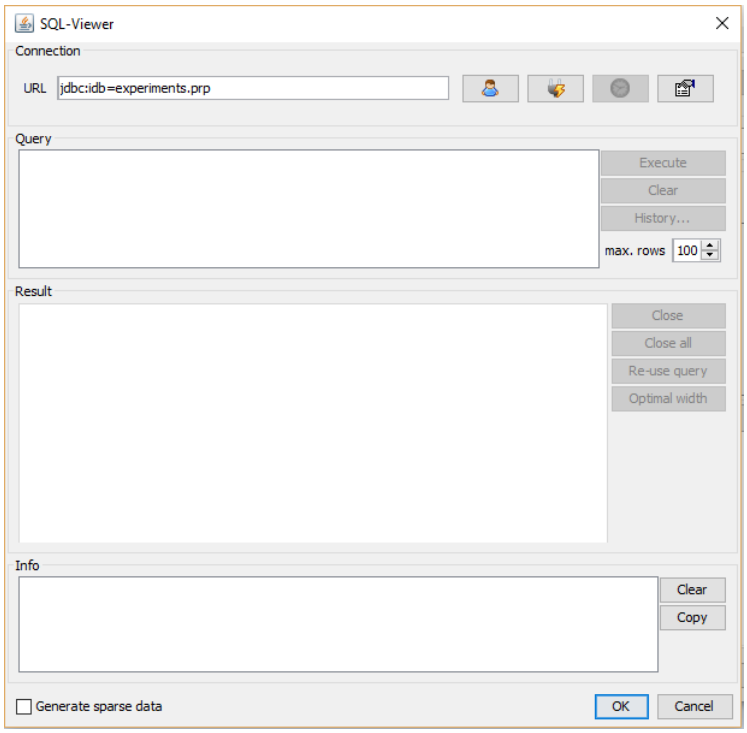
Figura 69 - Menu do Open URL.



Fonte: Próprio Autor

- **Open DB:** com este botão é possível obter os dados de um banco de dados. Seu menu de opções é ilustrado na Figura 70.

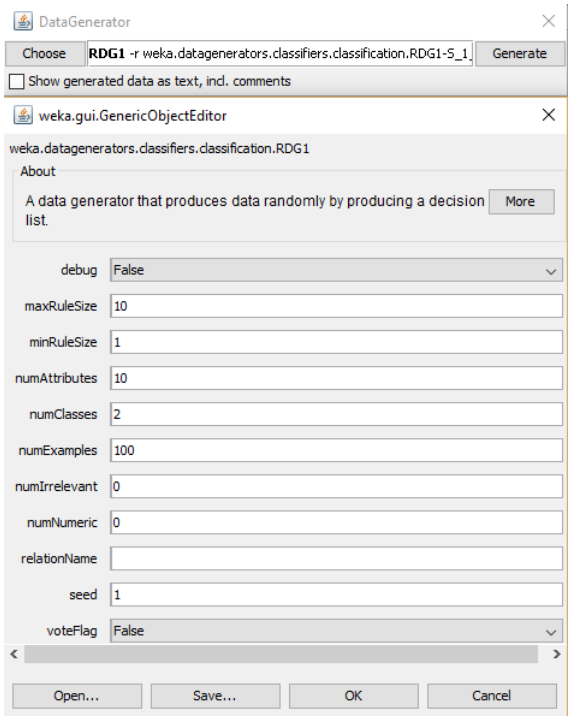
Figura 70 - Menu do Open DB.



Fonte: Próprio Autor

- **Generate:** é um gerador de dados que produz dados aleatórios através da produção de uma lista de decisão. A lista de decisão consiste em regras. Instâncias são geradas aleatoriamente uma a uma. Se a lista de decisão não classificar a instância atual, uma nova regra de acordo com esta instância é gerada e adicionada à lista de decisões. Seu menu de opções é ilustrado na Figura 71.

Figura 71 - Menu do Generate DB.



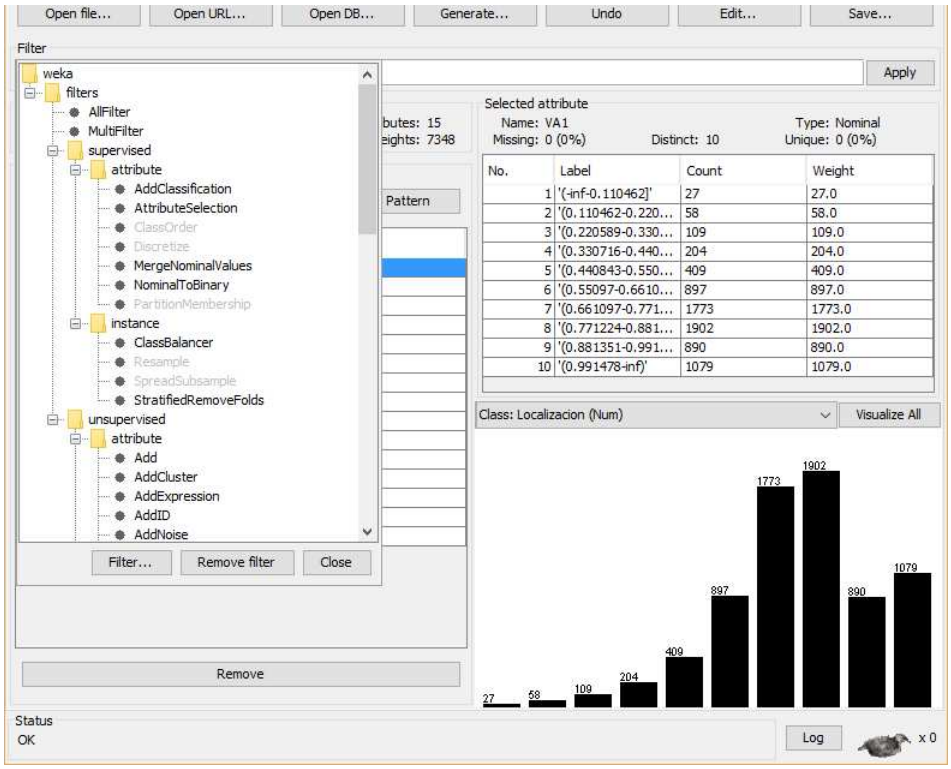
Fonte: Próprio Autor

Uma vez são escolhidos os dados que vão ser analisados, pode ser aplicado algum filtro ou realizar outras tarefas. Os botões **UNDO, EDIT** e **SAVE**, permitem realizar mudanças nos dados gerados no **Generate**.

A.2.2.2 Aplicação de Filtros

WEKA permite aplicar uma grande diversidade de filtros sobre os dados, com o objetivo de realizar transformações ao banco de dados. Ao ativar o botão **Choose** dentro da aba *Filter* é aberta uma lista de filtros que podem ser escolhidos considerando o tipo de algoritmo a ser utilizado, se é supervisionado ou não-supervisionado, além de considerar se o filtro vai ser aplicado aos atributos ou as instâncias. O menu de opções desta aba é ilustrado na Figura 72

Figura 72 - Aplicação de filtros da interface **Explorer**.

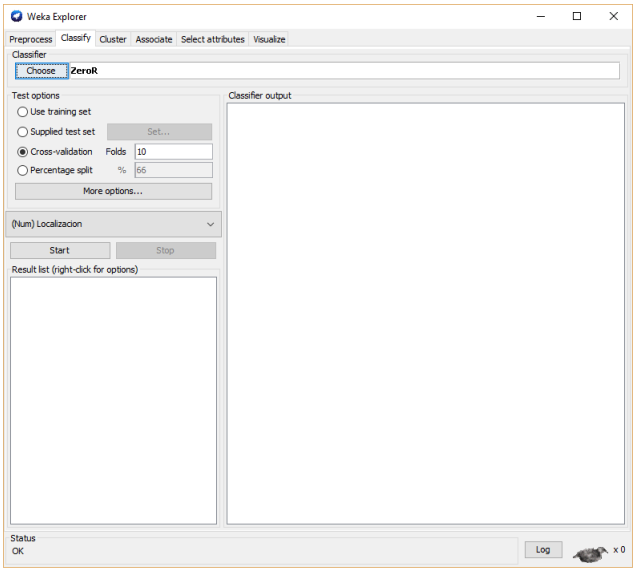


Fonte: Próprio Autor

A.2.2.3 Classify

Abrindo a segunda aba da parte superior da interface Explorer, é visualizada a tela no modo classificação. Nesta interface é possível realizar o processo de treinamento e validação dos algoritmos de aprendizado de dados utilizados na área de IC.

Figura 73 - Interface de classificação.



Fonte: Próprio Autor

Na hora de realizar o processo de classificação, primeiro deve ser escolhido o algoritmo de aprendizado de máquina que vai ser utilizado e posteriormente este tem que ser configurado, para isto, é pulsado o botão **Choose** dentro da área do *Classify*. Nesse momento uma lista de algoritmos se abre, permitindo selecionar o algoritmo desejado. Após a seleção do algoritmo, aparece uma etiqueta do lado do botão Choose, com a informação do filtro aplicado e a configuração interna dos parâmetros do algoritmo que foi escolhido.

Se for necessário realizar alguma alteração na configuração dos parâmetros do algoritmo, basta com dar um duplo clique sobre a etiqueta anteriormente mencionada, nesse momento aparece uma nova interface que permite modificar os valores definidos como padrão de cada um dos parâmetros de configuração do algoritmo.

Depois de ter escolhido e configurado o algoritmo de aprendizado de máquina que vai ser utilizado, o próximo passo é realizar a configuração do modo de treinamento, esta tarefa é realizada no **Test Options**. A ferramenta WEKA permite utilizar 4 modos de teste:

- **Use training set:** com esta opção o WEKA realiza o treinamento do algoritmo com todos os dados disponíveis e depois utiliza os mesmos dados para realizar o processo de validação.

- **Supplied test set:** escolhendo esta opção, é possível selecionar um banco de dados para realizar a validação do algoritmo que é treinado com dados repassados no início do processo. Para o carregamento dos dados de validação só é necessário usar o botão **Set**.

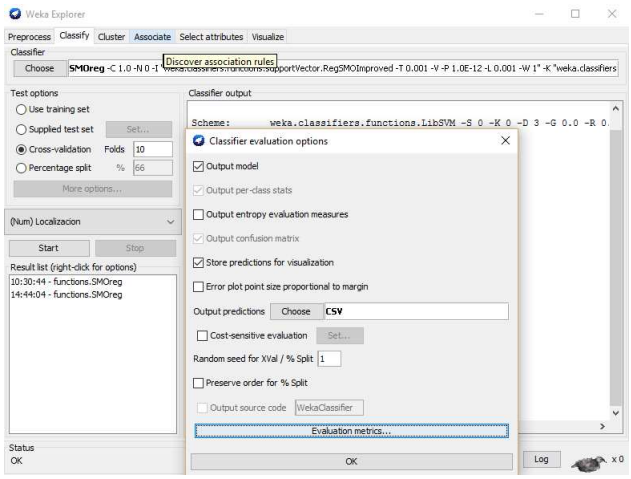
- **Cross validation:** com esta opção o WEKA realiza a validação cruzada que depende do

número de partições realizadas **(Folds)**. A validação cruzada consiste de um número *n* que é utilizado para realizar as divisões do banco de dados em *n* partes, onde cada uma destas partes é utilizada para realizar o treinamento do algoritmo em quanto outra parte é utilizada para realizar a validação do mesmo.

- **Percentage Split:** nesta opção, é definida a percentagem do banco de dados que é utilizada para realizar o treinamento do algoritmo de aprendizado de máquina, e com a percentagem restante e realizada a validação.

Após a definição do método de teste que será utilizado, o WEKA permite selecionar algumas outras opções utilizando o botão **More Options**, assim como ilustrado na Figura 74.

Figura 74 - Menu de opções adicionais na interface **Classify**.



Fonte: Próprio Autor

As opções adicionais são:

- **Output Model:** Permite visualizar o modelo construído.
- **Output per-class stats:** esta opção permite visualizar estatísticas referentes a cada atributo.
- **Output entropy evaluation mesures:** permite visualizar informações de medições da entropia na classificação.
- **Output confusion matrix:** permite visualizar a matriz de confusão do algoritmo utilizado. O número de colunas desta tabela é igual ao número de atributos que constam no banco de dados. A informação entregue pela matriz de confusão é de grande importância porque além de apresentar os erros produzidos, também permite visualizar o tipo de cada erro.

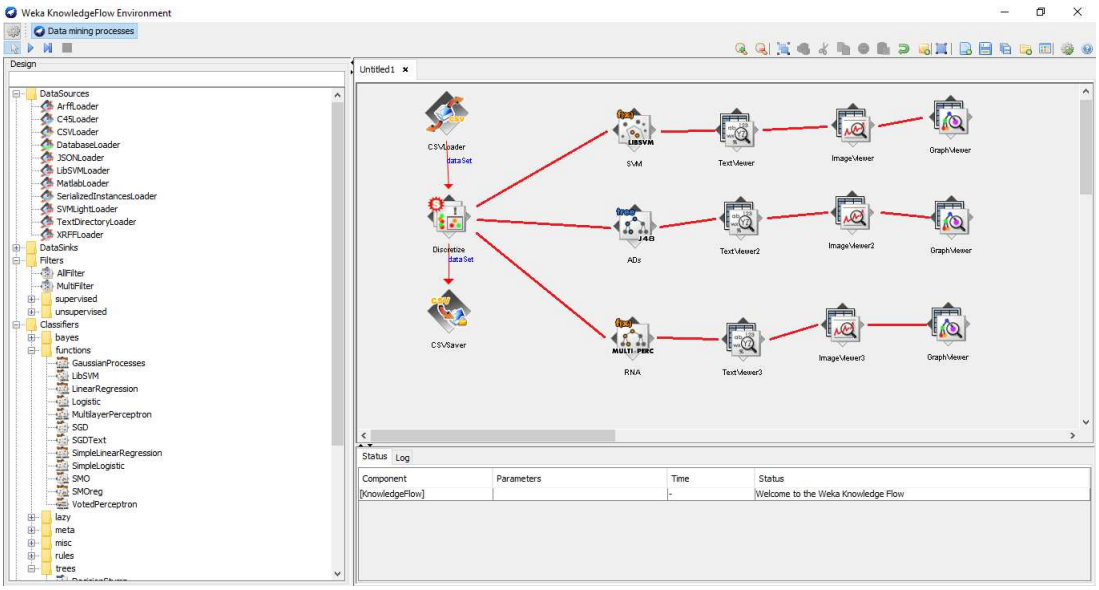
- **Store predictions for visualization:** conjunto de previsões na lista de resultados para posterior visualização
- **Error plot point size proportional to margin:** erros do algoritmo classificador, representa graficamente o tamanho do ponto que será definido proporcional ao valor absoluto da margem de predição.
- **Output predictions:** tipo de arquivo que será utilizado para salvar os resultados.
- **Cost-sensitive evaluation:** avalia os erros respeito à matriz de custo.
- **Preserve order for Split:** preserva a ordem em uma fração da percentagem.
- **Output source code:** permite visualizar a saída do classificador construído como código-fonte Java.

Para dar início ao treinamento e validação do algoritmo de aprendizado de máquina escolhido, só é necessário dar um click no botão *Start*. Uma vez inicializado o WEKA, na barra inferior pode ser visualizada a informação referente ao estado do teste. Quando o teste finaliza, o lado direito da interface é preenchido com a informação referente ao desenvolvimento do algoritmo e os resultados, tanto dos testes quanto da validação.

A.2.3 Knowledge Flow

Permite o desenvolvimento de projetos de IC em um ambiente gráfico com fluxos de informação, como ilustrado na Figura 75. Por outro lado, entre as muitas vantagens que possui, sobressai a possibilidade de realçar o layout intuitivo, e o fato de permitir o processamento de dados de forma incremental, o que lhe confere a possibilidade de aplicação a um conjunto de dados de elevada dimensão. Permite, também, o processamento paralelo, em que cada fluxo de dados distintos é processado no seu thread.

Figura 75 - Ambiente Knowledge Flow.

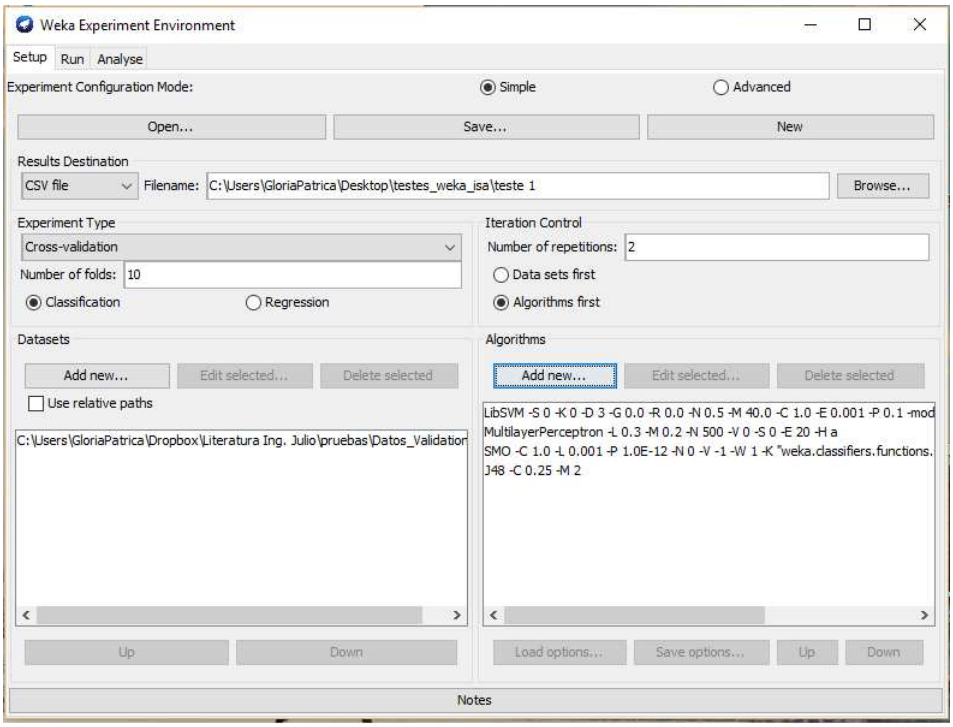


Fonte: Próprio Autor

A.2.4 Experimenter

Também tem um ambiente gráfico que permite testar técnicas diferentes em classificação ou regressão, permitindo compara-las. Embora também seja possível no *Explorer* e no *Knowledge Flow*, no *Experimenter* é possível escolher diversos conjuntos de dados a serem utilizados em uma experiência só, várias técnicas a serem experimentadas, o número de repetições (*runs*) do teste, entre outras escolhas, como ilustrado na Figura 76. Além do que, o teste pode ser executado sem ser necessária a supervisão do utilizador. Os resultados são guardados em ficheiros para posterior análise. É possível fazer a experiência com computação distribuída através de RMI (*Remote Method Invocation*). Esta é outra interface recomendada para realizar os testes, por esse motivo foi utilizada também para levar a cabo os testes e comparações entre os algoritmos utilizados nesta tese.

Figura 76 - Ambiente Experimenter.



Fonte: Próprio Autor

A ferramenta WEKA, reconhece os bancos de dados em vários formatos diferentes, tais como: .arff, .csv, .m, .xrff, .bsi, entre outros. Nos testes realizados ao longo desta pesquisa foi utilizado o formato .csv (comma-separated values) este é um formato que armazena dados tabelados, cujo grande uso data da época dos mainframes. Por serem bastante simples, arquivos .csv são comuns em todas as plataformas de computador. O csv é uma implementação particular de arquivos de texto separados por um delimitador, que usa a vírgula e a quebra de linha para separar os valores. Este formato também usa as aspas em campos no qual são usados os caracteres reservados. Essa robustez no formato torna o CSV mais atrativo que outros formatos digitais do mesmo segmento. Para formatar os dados como .csv, pode ser utilizado um simples editores de texto tais como são o notepad++ e o gedit, entre outros.

O WEKA disponibiliza uma ampla variedade de algoritmos de aprendizagem de máquina, entre os quais encontram-se os algoritmos de RNAs, MVSs, e as ADs. O algoritmo de RNAs utilizado nesta pesquisa desenvolve uma rede do tipo MultiLayer Perceptron (MLP). Este algoritmo pode ser utilizado para tarefas de regressão ou classificação. Os algoritmos que desenvolvem MVSs utilizam os métodos SMO (Support Vector Machine) que implementa o algoritmo de otimização mínima sequencia de (PLATT, 1998) , e LibSVM que é um pacote com implementações mais robustas e eficientes de diferentes SVM. Em quanto às ADs é utilizado o J48 para classificação, que não é mais do que uma implementação do conhecido algoritmo C4.5 (desenvolvido por J. Quinlan) (QUINLAN, 1996).

A.2.5 Classificação com RNAs e MVSs no WEKA

Como mencionado anteriormente, o *Explorer* e o *Experimenter* são as interface mais adequadas para realizar os testes, já que, é possível escolher várias tarefas e técnicas na hora de realizar o treinamento e validação dos algoritmos de IC. Além disso, os testes podem ser executados sem ser necessária a intervenção do utilizador, tendo posteriormente acesso aos resultados armazenados num arquivo digital. Rassaltando que é possível fazer os testes utilizando computação distribuída, o que pode acelerar consideravelmente os tempos dos testes. A possibilidade de fazer os testes utilizando computação distribuída é muito importante, pois um elevado número de conjunto de dados, conjugado com as várias iterações e técnicas requer de um elevado poder computacional.

O primeiro passo para iniciar o trabalho com a interface *Explorer* é definir a origem dos dados. Os testes iniciam no painel *Preprocess*, a primeira ação a ser realizada é escolher a base de dados que vai ser analisada, uma vez realizada esta ação é aplicado um filtro aos dados, este filtro permite modificar de certo modo os dados, por exemplo adicionando ruído aos dados. Neste módulo é possível selecionar conjunto de dados em diversos formatos, excluir atributos e aplicar filtros aos dados.

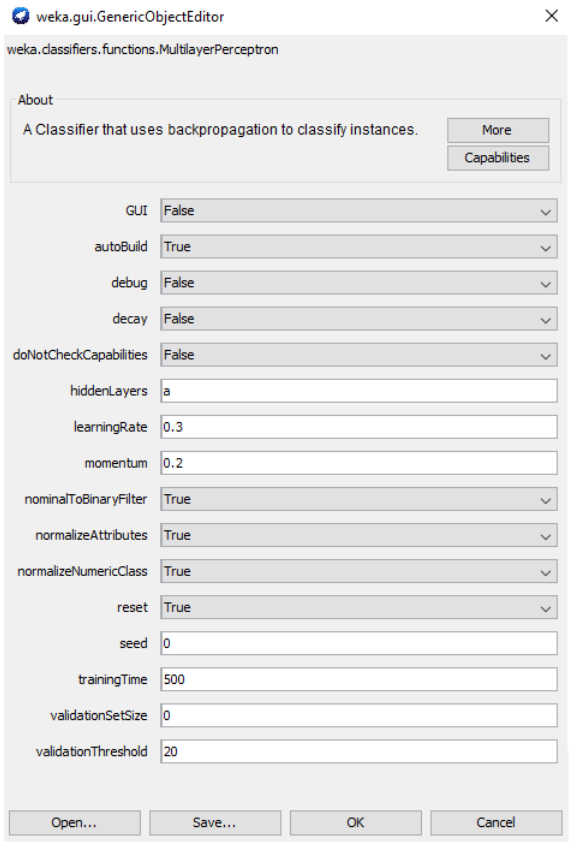
O módulo de *Classify* permite treinar e testar sistemas de aprendizagem que classificam ou realizam uma regressão dos dados selecionados em *Preprocess*. Neste módulo é possível selecionar e configurar diversos classificadores, entre eles se encontram os utilizados nesta pesquisa, RNAs, MVSs e ADs. Também permite fornecer arquivos de teste e escolher a metodologia de teste entre as seguintes opções:

- **Use training set:** usa os casos de treino como de teste;
- **Supplied test set:** permite selecionar um arquivo com os casos de teste;
- **Cross-validation:** usa validação cruzada do tipo k-fold;
- **Percentage Split**: usa uma certa percentagem dos dados para teste.

Finalmente é escolhida a técnica de IC (RNAs, MVSs e ADs), que é utilizada tanto nos testes como na validação. Uma vez escolhido o algoritmo de IC, são modificados seus parâmetros até chegar a obter um valor de resposta aceitável.

Na Figura 77 pode ser vista a lista de parâmetros que tem que ser calibrados para o algoritmo de RNAs MLP até obter o resultado esperado.

Figura 77 - Parâmetros a serem calibrados de uma RNA MLP.



Fonte: Próprio Autor

A Tabela 26 apresenta os nomes dos parâmetros associados a técnica de RNA, a descrição dos mesmos e o valor padrão que é utilizado.

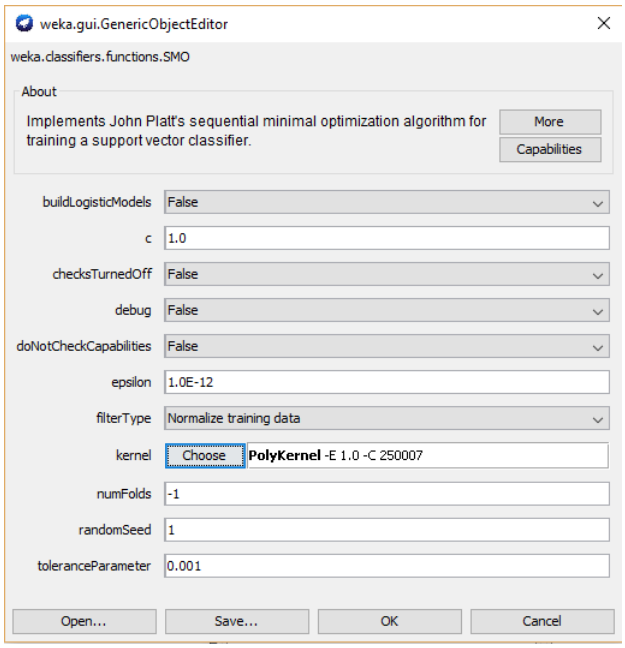
Na Figura 78 pode ser vista a lista de parâmetros que têm que ser calibrados do algoritmo de MVS SMO, até obter o resultado esperados.

Tabela 26 - Descrição dos parâmetros associados a técnica de RNA.

Parâmetro	Descrição	Valor Padrão
GUI	Permite visualizar uma interface GUI para configuração da topologia da rede.	FALSO
AutoBuild	Constrói as camadas intermédias da rede.	VERDADEIRO
Debug	Apresentação de informação adicional.	FALSO
Decay	Diminui a taxa de aprendizagem: a taxa de aprendizagem de cada iteração é obtida dividindo-a pelo número da iteração.	FALSO
DoNotCheckCapabilities	Se definido, as capacidades dos classificadores não são verificados antes do classificador ser construído (Use com cuidado para reduzir o tempo de execução).	FALSO
HiddenLayers	Define as camadas intermédias.	(Atributos+classes)/2
LearningRate	Taxa de aprendizagem.	0,3
Momentum	Taxa de aprendizagem.	0,2
NominalToBinaryFilter	Converte os atributos nominais em binários.	VERDADEIRO
NormalizeAttributes	Normaliza os atributos numéricos.	VERDADEIRO
NormalizeNumericClass	Normaliza o atribuído a prever caso seja numérico.	VERDADEIRO
Reset	Permite que se reinicie a aprendizagem com uma taxa de aprendizagem menor caso o algoritmo esteja a divergir.	VERDADEIRO
Seed	Semente *(seed)* para a geração aleatória dos pesos iniciais das sinapses.	0
TrainingTime	Número de iterações de treino.	500
ValidationSetSize	Percentagem dos dados a serem utilizados para validação.	0
ValidationThreshold	Número de vezes que o erro pode piorar nos dados de validação até terminar o treino.	20

Fonte: Próprio Autor

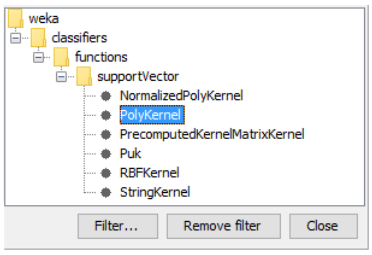
Figura 78 - Parâmetros a serem calibrados de um algoritmo de MVS SMO.



Fonte: Próprio Autor

Deve-se considerar que os algoritmos de MVS, não só dependem dos parâmetros apresentados na Figura 78, também é necessário escolher o kernel que vai ser utilizado para realizar o processo de aprendizagem, na Figura 79 são apresentadas as opções de kernel.

Figura 79 - Parâmetros a serem calibrados de um algoritmo de MVS SMO.



Fonte: Próprio Autor

A Tabela 27 apresenta os nomes dos parâmetros associados a técnica de MVS, a descrição dos mesmos e o valor padrão que são utilizados.

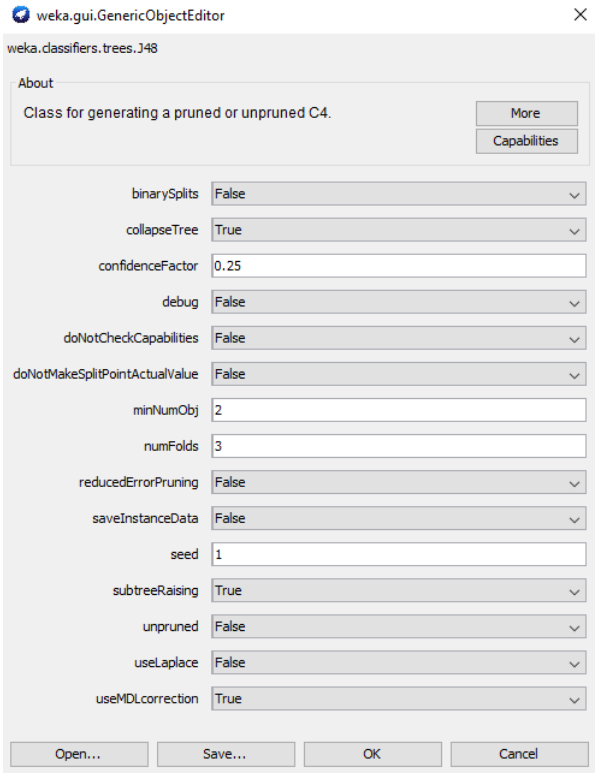
Na Figura 80 pode ser vista a lista de parâmetros que têm que ser calibrados do algoritmo de AD J48, até obter os resultados esperados.

Tabela 27 - Descrição dos parâmetros associados a técnica de MVS.

Parâmetro	Descrição	Valor Padrão
buildLogisticModels	Permite modelação logística.	FALSO
c	Parâmetro C de complexidade.	1.0
checksTurnedOff	Realiza verificações de demora, use com cautela.	FALSO
Debug	Apresentação de informação adicional.	FALSO
DoNotCheckCapabilities	Se definido, as capacidades dos classificadores não são verificadas antes do classificador ser construído (Use com cuidado para reduzir o tempo de execução).	FALSO
epsilon	O valor do epsilon de erro de arredondamento (não deve ser alterado).	1.0E-12
FilterType	Determina a transformação dos dados.	*Normalize Training Data*
Kernel	O Kernel a ser usado.	Polykernel
numFolds	O número de desdobramentos para validação cruzada usado para gerar dados de treinamento para modelos logísticos (-1 significa que os dados de treinamento são utilizados).	-1
randomSeed	Semente *(seed)* para geração aleatória de elementos durante o CV.	1
toleranceParameter	Tolerância (não deve ser alterado).	0.001

Fonte: Próprio Autor

Figura 80 - Parâmetros a serem calibrados de um algoritmo de AD J48.



Fonte: Próprio Autor

A Tabela 28 apresenta os nomes dos parâmetros associados a técnica de AD, a descrição dos mesmos e o valor padrão que é utilizado.

Tabela 28 - Descrição dos parâmetros associados ao algoritmo J48.

Parâmetro	Descrição	Valor Padrão
binarySplits	Divisão binária em atributos nominais.	FALSO
CollapseTree	Removidas as peças que não reduzem o erro de treinamento.	VERDADEIRO
ConfidenceFactor	Factor de confiança utilizado na poda.	0,25
Debug	Apresentação de informação adicional.	FALSO
DoNotCheckCapabilities	Se definido, as capacidades dos classificadores não são verificadas antes do classificador ser construído (Use com cuidado para reduzir o tempo de execução).	FALSO
DoNotMakeSplitPointActual Value	Se for verdadeiro, o ponto de divisão não é atribuída a um valor de dados reais. Isso pode produzir aceleração substancial para grandes conjuntos de dados com atributos numéricos.	FALSO
MinNumObj	Número mínimo de instâncias por folha.	2
numFolds	Determina a quantidade de dados utilizados para a poda de erros reduzida. Uma dobra é usada para a poda, o resto para o crescimento da árvore.	3
ReducedErrorPruning	Permite optar por *reduced error pruning* ou poda do C.4.5.	FALSO
SaveInstanceData	Opção para guardar os dados de treinamento para visualização.	FALSO
Seed	Semente (*seed*) para gerar aleatoriamente os índices quando se usa *ReducedErrror Pruning*.	1
subtreeRaising	Se for considerada, permite a *subtree raisin* na poda.	VERDADEIRO
Unpruned	Impede a poda.	FALSO
UseLaplace	Método Laplace na contagem das folhas.	FALSO
UseMDLcorrection	Correção MDL é usado quando encontrar divisões nos atributos numéricos.	VERDADEIRO

Fonte: Próprio Autor

Como foi visto nas figuras e tabelas anteriores, cada um dos algoritmos utilizados para realizar o aprendizado de máquina, tem relacionados vários parâmetros que devem ser calibrados até chegar a um resultado aceitável. Cada teste foi realizado 30 vezes por cada parâmetro que era mudado. Uma vez que era alcançado um valor aceitável para um parâmetro, eram modificados outros parâmetros. Os testes iniciais foram realizados com os valores padrões do WEKA. No final de cada teste, são analisados os resultados e são armazenados em um banco de dados para posteriormente serem comparados com outros resultados do mesmo algoritmo e dos outros algoritmos.