UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
ESCOLA DE ENGENHARIA
DEPARTAMENTO DE ENGENHARIA ELÉTRICA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

RODRIGO DANIEL TREVIZAN

DETECÇÃO E IDENTIFICAÇÃO DE PERDAS COMERCIAIS EM SISTEMAS DE DISTRIBUIÇÃO: METODOLOGIA BASEADA EM FLORESTA DE CAMINHOS ÓTIMOS

Porto Alegre
2014

RODRIGO DANIEL TREVIZAN

DETECÇÃO E IDENTIFICAÇÃO DE PERDAS COMERCIAIS EM SISTEMAS DE DISTRIBUIÇÃO: METODOLOGIA BASEADA EM FLORESTA DE CAMINHOS ÓTIMOS

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal do Rio Grande do Sul como parte dos requisitos para a obtenção do título de Mestre em Engenharia Elétrica.
Área de concentração: Energia

ORIENTADOR: Prof. Dr. Arturo Suman Bretas

Porto Alegre
2014

RODRIGO DANIEL TREVIZAN

DETECÇÃO E IDENTIFICAÇÃO DE PERDAS COMERCIAIS EM SISTEMAS DE DISTRIBUIÇÃO: METODOLOGIA BASEADA EM FLORESTA DE CAMINHOS ÓTIMOS

Esta dissertação foi julgada adequada para a obtenção do título de Mestre em Engenharia Elétrica e aprovada em sua forma final pelo Orientador e pela Banca Examinadora.

Orientador: _____
Prof. Dr. Arturo Suman Bretas, UFRGS
Doutor pela Virginia Polytechnic and State University – Blacksburg, Estados Unidos

Banca Examinadora:

Prof. Dr. Roberto Chouhy Leborgne, UFRGS – PPGEE
Doutor pela Chalmers University of Technology – Gotemburgo, Suécia

Prof. Dr. Gustavo Dorneles Ferreira, UFRGS – PPGEE
Doutor pela Universidade Federal do Rio Grande do Sul – Porto Alegre, Brasil

Prof. Dr. Daniel da Silva Gazzana, UFRGS – PPGEE
Doutor pela Universidade Federal do Rio Grande do Sul – Porto Alegre, Brasil

Coordenador do PPGEE: _____
Prof. Dr. Alexandre Sanfelice Bazanella

Porto Alegre, dezembro de 2014.

AGRADECIMENTOS

Aos meus pais Hertes e Mirian, que são os maiores responsáveis por eu ter conseguido trilhar este caminho graças aos ensinamentos, ao amor, ao carinho, ao exemplo, aos valores, às renúncias e ao suporte que sempre proporcionaram.

Ao meu irmão Ricardo por todo apoio e por ser um grande amigo e companheiro.

Aos colegas do LASEP pela amizade, pelo ambiente acolhedor, pelos momentos inesquecíveis e pelo companheirismo durante este mestrado. Em especial ao colega Aquiles que auxiliou muito na pesquisa.

Aos demais colegas do PPGEE pelas experiências que dividimos e pelo ambiente diverso e enriquecedor.

Ao Prof. Arturo Bretas por ter me aceitado como seu orientando, pela paciência, pelos ensinamentos, pelo exemplo, pela orientação e por mostrar uma nova visão da vida acadêmica.

Aos professores Daniel da Silva Gazzana, Gustavo Dorneles Ferreira e Roberto Chouhy Leborgne por aceitarem o convite de participar da minha banca e pelo tempo dedicado à avaliação do meu trabalho.

Aos demais professores do Programa de Pós-Graduação em Engenharia Elétrica, em especial aos do LASEP, pelos conhecimento e experiências compartilhadas em sala de aula.

Aos servidores do PPGEE pela cordialidade, pelos bons embora breves momentos compartilhados, pela solicitude e pelo importante trabalho desempenhado no programa.

À Universidade Federal do Rio Grande do Sul pela infra-estrutura disponibilizada, pelo conhecimento e experiências proporcionados, que transformaram minha vida.

À CAPES por ter financiado o meu trabalho, tornando a sua realização possível.

Às empresas Neo Domino Consultoria e Pesquisa Ltda. (NEO DOMINO), CHESP, CERRP, CERPRO, CERNHE, CERIPA, CERAL-DIS, CETRIL, CERIM, CERMC, CERIS, CEDRI, CERES, CEDRAP, ELFSM, EFLJC, COOPERALIANÇA e CERCOS pelo suporte para o desenvolvimento deste trabalho mediante a realização de um projeto de P& D para a ANEEL (projeto 0103-0002/2011).

Ao povo brasileiro, cujo trabalho mantém as instituições em que tive o privilégio de receber minha educação superior e ao qual serei eternamente grato.

RESUMO

O sistema elétrico brasileiro possui atualmente níveis de perdas elétricas da ordem de 15%. Destes, aproximadamente a metade são provenientes das chamadas perdas comerciais (PC) que ocorrem nos sistemas de distribuição. As PC são a soma de toda energia não faturada pelas distribuidoras, á exceção das perdas técnicas. As suas causas mais frequentes são os furtos de energia elétrica, fraudes e defeitos em medidores. Os custos provenientes dessas perdas são normalmente repassados pelas distribuidoras aos consumidores regulares. No entanto, novas regulamentações do regulador brasileiro do sistema elétrico, a Agência Nacional de Energia Elétrica (ANEEL), impõem limite a esse repasse, o que cria nas concessionárias um maior incentivo para o seu combate. Entre as metodologias empregadas para a mitigação de PC, tem sido destacadas na literatura aquelas baseadas na análise das bases de dados de clientes das empresas distribuidoras com o objetivo de reconhecer padrões de clientes irregulares. Neste contexto, neste trabalho é proposto e desenvolvido um sistema de combate a PC baseado no classificador supervisionado Floresta de Caminhos Ótimos (*Optimum-Path Forest*, OPF). São propostas a utilização de dados categóricos e a normalização de dados como modificações nos métodos encontrados na literatura. Os testes com o sistema desenvolvido são aplicado a uma base de dados sintetizada a partir de clientes residenciais, diferentemente de trabalhos em que se utilizaram dados de consumidores comerciais e industriais. Os resultados mostram que as modificações propostas podem melhorar o desempenho do OPF. O comparativo com outros métodos de classificação reafirma a eficiência do OPF mas contesta alguns resultados presentes na literatura.

Palavras-chave: Perdas Comerciais, Perdas em Sistemas de Distribuição, Perdas Não-Técnicas, Reconhecimento de Padrões, Aprendizado de Máquina, Floresta de Caminhos Ótimos.

ABSTRACT

The Brazilian electric power system has about 15% of losses. About a half of this amount is due to the so called commercial losses. The commercial losses are the sum of the unbilled energy less the technical losses. The commercial losses are mainly caused by electricity theft, frauds in electricity meters and electricity meter failure. The financial costs caused by these losses are included in the electricity bill, paid by the regular consumers. New regulations approved by the Brazilian regulatory agency for the electric system create a limit for this, which stimulates the investments in commercial loss mitigation by distribution companies. Among the methods used to mitigate commercial losses, those based on pattern recognition of irregular consumers within electric companies' clients' databases are some of the most promising. In this work, a system for commercial losses mitigation based on the supervised classifier Optimum-Path Forest (OPF) is studied and developed. Categorical data and data normalization are proposed as methods for improving classifier performance. In order to check the system performance, tests are conducted on a database derived from residential consumer data found in the literature, differently from other works which proposed data classification for commercial and industrial consumers only. The results show that using categorical data and normalization may improve OPF performance. Comparing this method with other classifiers confirms OPF's efficiency but contests some results shown in the literature.

Keywords: Commercial Losses, Losses in Distribution Systems, Non-Technical Losses, Pattern Recognition, Machine Learning, Optimum-Path Forest.

SUMÁRIO

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS

ANEEL	Agência Nacional de Energia Elétrica
AMI	*Advanced Metering Infrastructure*
ANN	*Artificial Neural Network*
DC	Dados Categóricos
EPE	Empresa de Pesquisa Energética
F	F-score
IA	Inteligência Artificial
LDA	*Linear Discriminant Analysis*
MLP	*Multi-Layer Perceptron*
MST	*Minimum Spanning Tree*
OPF	*Optimum-Path Forest*
PC	Perdas Comerciais
PEE	Programa de Eficiência Energética
PNT	Perdas Não-Técnicas
PRODIST	Procedimentos de Distribuição
PRORET	Procedimentos de Regulação Tarifária
PT	Perdas Técnicas
RBF	*Radial Basis Function*
RNA	Rede Neural Artificial
SENS	Sensibilidade
SVM	*Support Vector Machine*
SVM-Linear	*Support Vector Machine using Linear kernel*
SVM-RBF	*Support Vector Machine using Radial Basis Function kernel*
TCC	Total de classificação correta
TD	Transformador de Distribuição
T&D	Transmissão e Distribuição

UC	Unidade Consumidora
VPP	Valor preditivo positivo

LISTA DE SÍMBOLOS

Li	Indicador Precisão
FP	Falsos positivos
FN	Falsos negativos
VP	Verdadeiros positivos
VN	Verdadeiros negativos
s	Amostra
t	Amostra
Z_1	Conjunto de treinamento
Z_2	Conjunto de aprendizado
Z_3	Conjunto de teste
S	Conjunto de protótipos

1 INTRODUÇÃO

1.1 Motivação

O preço da energia elétrica é um fator de grande contribuição nos resultados macroeconômicos de um país. O aumento das tarifas de energia elétrica causa onerações do setor produtivo ao consumidor residencial. No Brasil, o aumento nacional das tarifas que tem sido aprovado está tendo grandes impactos econômicos, inclusive contribuindo para o aumento da inflação no país (INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA, 2014). Este cenário reforça a importância de se estudar maneiras de reduzir os custos da energia elétrica, de modo a promover a modicidade tarifária.

O consumo de energia elétrica no Brasil tem expandido ano após ano a taxas acima do crescimento populacional. Entre os anos de 1970 e 2005, o consumo de energia elétrica cresceu em média 6,7% ao ano (EMPRESA DE PESQUISA ENERGÉTICA, 2007, p. 189) enquanto no mesmo período o crescimento populacional foi de 1,46% ao ano (EMPRESA DE PESQUISA ENERGÉTICA, 2007, p. 15). As projeções para os anos de 2013 a 2022 indicam crescimento populacional de 0,6% ao ano e de demanda por energia elétrica de 4,7%, acima do aumento de consumo final energético estimado em 4,5% (EMPRESA DE PESQUISA ENERGÉTICA, 2013, p. 38). Essa tendência de aumento no consumo de energia em todas as suas formas, em especial de energia elétrica, contrasta com tendências mundiais de diminuição de emissão de gases de efeito estufa, provenientes da queima de combustíveis fósseis, e com as crescentes exigências ambientais para implantações de empreendimentos de grandes proporções para geração de energia elétrica. Somado a isso, o esgotamento do potencial hidrelétrico nas regiões próximas aos grandes centros consumidores do Brasil leva à expansão prevista de unidades geradoras hidrelétricas na região Norte (EMPRESA DE PESQUISA ENERGÉTICA, 2013, p. 87), responsável por apenas 7,2% da demanda nacional, o que é indesejável do ponto de vista ambiental pelas características das hidrelétricas que são construídas neste região, cria desafios técnicos por conta da necessidade de grandes linhas de transmissão e se torna portanto economicamente menos vantajoso do que usinas localizadas perto dos grandes centros consumidores. Para atingir o balanço energético, ainda será necessário aumentar a participação de fontes de energia cujo custo de geração de energia elétrica é maior do que as hidrelétricas, tais como usinas eólicas, de biomassa (EMPRESA DE PESQUISA ENERGÉTICA, 2013, p. 99) e solar (EMPRESA DE PESQUISA ENERGÉTICA, 2013, p. 339). Logo, pode-se projetar cenários futuros com tendências para o aumento das tarifas de energia elétrica.

O investimento em eficiência energética tem sido apontado como uma das saídas para a redução dos investimentos em infraestrutura no setor elétrico. O Plano Decenal de Expansão de Energia 2022 projeta conservação de 5,8% de energia elétrica até o ano de

2022 (EMPRESA DE PESQUISA ENERGÉTICA, 2013, p. 340). Para que esses números sejam alcançados, é preciso que os progressos advindos da tendência da utilização de tecnologias mais eficientes em novos produtos elétricos e eletrônicos seja somada a políticas públicas para a promoção da eficiência energética. Algumas dessas políticas são promovidas pela Agência Nacional de Energia Elétrica (ANEEL). Por meio de chamadas públicas, a ANEEL incentiva iniciativas para melhoria de eficiência energética por meio do Programa de Eficiência Energética (PEE) e do Programa de Pesquisa e Desenvolvimento Tecnológico do Setor de Energia Elétrica. Estes projetos são orientados a temas estratégicos ou prioritários para o setor elétrico nacional, sob critério definido pela agência.

Do ponto de vista das distribuidoras de energia elétrica, promover eficiência energética também passa por reduzir as perdas elétricas na distribuição. As perdas em sistemas elétricos podem ser classificadas em duas categorias: técnicas e não-técnicas. As perdas técnicas (PT) são aquelas inerentes ao transporte da energia elétrica. Já as perdas não-técnicas (PNT) são caracterizadas por energia consumida mas não faturada pela distribuidora, gerando perdas econômicas. Por esse motivo, neste trabalho, será utilizada a terminologia perdas comerciais (PC) como sinônimo de PNT. A origem mais comum desse tipo de perdas são defeitos em medidores e consumidores irregulares.

As perdas em sistemas elétricos de potência são um problema de ordem mundial. Como é mostrado na tabela 1, entre os anos 1980 e 2000 houve um aumento na quantidade de perdas nos sistemas de transmissão e distribuição (T&D) em todo o mundo. Pode-se perceber que há uma correlação entre o alto nível de perdas e a localização dos países em desenvolvimento.

Tabela 1: Perdas em T&D no mundo por região.

Região	**Países**	**Perdas em T&D [%]**		**Mudança [%]**
		1980	**2000**	
Europa Ocidental	17	7,71	7,56	-0,15
Europa Oriental	24	9,68	18,18	8,5
Oriente Médio e Norte da África	11	11,18	19,63	8,45
África	11	14,6	19,95	5,35
América do Norte	3	9,67	9,38	-0,29
América do Sul	9	13	17,23	4,23
América Central e Caribe	9	15,5	21,68	6,18
Sul da Ásia	5	25,2	27,55	2,35
Sudeste da Ásia	7	12,14	13,32	1,18
Leste da Ásia e Australásia	6	8,67	7,65	-1,02
Total	**102**	**11,69**	**16,22**	**4,54**

Fonte: (SMITH, 2004).

As perdas globais anuais de energia elétrica no Brasil são da ordem de 52 TWh, representando aproximadamente 15 % da energia total (VIDINICH; NERY, 2009). Fazendo uma comparação, a produção anual da usina hidrelétrica de Belo Monte é de aproximadamente 40 TWh (NORTE ENERGIA, 2011). As PC representam 44 % do total de perdas e o seu custo anual para a sociedade brasileira é de aproximadamente R$5,5 bilhões, chegando a R$7,3 bilhões se forem levados em conta tributos que deixam de ser arrecadados

como ICMS, PIS e COFINS (VIDINICH; NERY, 2009). A eficiência energética do sistema de distribuição brasileiro quando comparada àquela de outros países é considerada baixa, como mostram os dados da tabela 2. Quando comparado a países desenvolvidos como EUA ou Suécia ou a países em que há grandes investimentos em infra-estrutura como a China, as perdas são aproximadamente da ordem de grandeza da metade daquelas que ocorrem no distribuição no Brasil. Países de extensão territorial superior àquela do Brasil, como a Rússia, também possuem níveis de perdas inferiores. Logo, pode-se concluir que reduzir o nível de perdas é possível e que depende de investimentos para que ocorra.

Tabela 2: Comparação entre perdas na distribuição em diversos países.

País	**Perdas (2011)**
África do Sul	9,20%
Argentina	14,10%
Brasil	**15,50%**
China	6,00%
EUA	6,20%
Índia	22,70%
Rússia	10,80%
Suécia	7,60%

Fonte: (U. S. ENERGY INFORMATION ADMINISTRATION, 2012).

A redução de PT pode ser alcançada pela substituição de componentes da infra-estrutura por outros mais eficientes, sendo portanto dependente da tecnologia utilizada. Outra maneira de mitigar essas perdas é otimização da alocação de equipamentos e cargas nos sistemas elétricos.

Neste trabalho, o foco será na redução de PC. O combate às PNT é realizado geralmente de duas formas. Uma delas é a implementação de ações que visam dificultar as fraudes nos medidores que apuram o consumo de energia nas unidades consumidoras (UC) e o furto de energia. Outra maneira é a eliminação das fontes de PC por meio da regularização das UC. Esta segunda opção inclui auditorias dos medidores realizadas em campo por equipes especializadas em encontrar irregularidades em UC. No entanto, essas inspeções exigem equipes treinadas e equipadas para o sucesso da auditoria, o que torna essa tarefa custosa. Além disso, as distribuidoras de energia elétrica possuem de dezenas a centenas de milhares de UC, o que torna impraticável a realização de inspeções por varredura em todos os consumidores. Somado a isso tudo, a realização de inspeções em UC acarreta em custo não mensurável às distribuidoras por conta de possíveis constrangimentos causados aos clientes.

A resolução do problema de detecção e localização de PC passa pela utilização de métodos que possam reduzir o universo de clientes a serem inspecionados a uma lista cuja quantidade de elementos seja pequena a ponto de tornar sua inspeção possível. Além disso, deseja-se que as inspeções realizadas nos elementos da lista gerada possua alta taxa de identificação de irregularidades. Desta maneira, se torna possível reduzir os custos financeiros das distribuidoras ao mesmo tempo em que se evitam os possíveis danos aos consumidores inspecionados.

A mitigação de PC é um dos temas estratégicos contemplados nas chamadas públicas para projetos de pesquisa e desenvolvimento promovidos pela ANEEL. Tendo em vista

o contexto aqui descrito, fica claro que é muito importante o combate às PNT para garantir a modicidade tarifária e mitigar a oneração causada por essas perdas à sociedade brasileira. Além do incentivo a soluções inovadoras para mitigação de PNT, a ANEEL pode penalizar distribuidoras que possuam altos níveis de perdas. As perdas são repassadas em parte aos clientes das distribuidoras, e a quantidade máxima desse prejuízo que se permite embutir na fatura das unidades consumidoras é definida pela própria agência com base em metodologias específicas que levam em conta as características de cada distribuidora (AGÊNCIA NACIONAL DE ENERGIA ELÉTRICA (Brasil), 2014). Buscando atender as exigências das agências reguladoras, as empresas de energia elétrica têm destinado investimentos com o objetivo de aumentar o seu desempenho técnico e financeiro, obtendo maior produtividade, eficiência e rentabilidade. Levando em consideração o que foi exposto até aqui, pode-se afirmar que o tema da eficiência técnica e financeira das empresas distribuidoras de energia elétrica é de grande importância para a sociedade brasileira.

1.2 Objetivos

O objetivo principal deste trabalho é o desenvolvimento de um método para a detecção e identificação de PC em sistemas elétricos de distribuição. Este método deve gerar uma lista de UC suspeitas em que haja alta probabilidade de se confirmar a existência de PNT. A relação de elementos suspeitos serviria como guia para equipes de inspeção de empresas distribuidoras de energia elétrica realizem a auditora de cada caso *in loco*, desta forma contribuindo para o combate a PC em sistemas elétricos de distribuição.

É desejável que a solução deste problema seja inovadora e que o seu desempenho seja superior ao já encontrado na literatura. Também é incluída como objetivo a revisão da literatura, para identificar outros métodos que possam ser aplicados ao problema da detecção e identificação de perdas comerciais em sistemas de distribuição. Por fim, para avaliar o desempenho do método desenvolvido, é necessária a implementação de um sistema computacional capaz de avaliar um caso real de uma empresa distribuidora.

1.3 Estrutura do Trabalho

Este trabalho é composto por 7 capítulos. No capítulo 2 são fornecidos maiores detalhes sobre as perdas nos sistemas de distribuição com enfoque nas particularidades do sistema elétrico brasileiro. As causas de PT e PC são apresentadas, bem como o seu processo de apuração. No capítulo 3 são apresentados alguns métodos empregados para mitigação de PC encontrados na literatura. Ao final do capítulo são feitas comparações entre os métodos mencionados, justificando a escolha feita pelo OPF. No capítulo 4 o classificador OPF é apresentado de forma detalhada, com os seus principais algoritmos. A estrutura da metodologia de localização e identificação de PC é apresentada no capítulo 5. Alguns estudos de caso são realizados para apuração do desempenho do sistema desenvolvido 6. Por fim, uma conclusão é apresentada no capítulo 7, juntamente com os trabalhos futuros a serem desenvolvidos.

2 PERDAS EM SISTEMAS ELÉTRICOS DE POTÊNCIA

2.1 Introdução

Antes de estudar métodos para redução de perdas em sistemas de distribuição, é necessário entender quais são as suas causas. Apesar de possuírem causas distintas, as estimativas de PT e PC são influenciadas pelo modo que são obtidas. Neste capítulo são apresentados os conceitos de perdas técnicas e não-técnicas, juntamente com as suas causas. Para explicar o problema da estimação de PC, os métodos de cálculo segundo a definição da ANEEL também são apresentados.

2.2 Definições

As perdas nos sistemas elétricos de potência são definidas pelo PRODIST, no módulo 7 (AGÊNCIA NACIONAL DE ENERGIA ELÉTRICA, 2013a). Neste módulo alguns indicadores de perdas são assim definidos:

- Energia Fornecida (EF): energia ativa efetivamente entregue e medida, ou estimada, nos casos previstos pela legislação, às unidades consumidoras, outras distribuidoras e consumidores livres, mais o consumo próprio, em megawatt-hora (MWh);
- Energia Injetada (EI): energia ativa efetivamente recebida e medida de um agente, em megawatt-hora (MWh);
- Perdas Técnicas do Segmento (PTS) (i): perdas técnicas para cada segmento, em megawatt-hora (MWh);
- Perdas Técnicas (PT): corresponde à soma das perdas técnicas de todos os segmentos, em megawatt-hora (MWh);
- Perdas na Distribuição (PD): corresponde à diferença entre EI e EF, em megawatt-hora (MWh);
- Perdas Não Técnicas (PNT): corresponde à diferença entre PD e PT, em megawatt-hora (MWh) ;

Matematicamente, para um sistema com n segmentos isso é definido pelas equações (1) a (3).

$$PD = EI - EF \tag{1}$$

$$PT = \sum_{i=1}^{n} PTS_i \tag{2}$$

$$PNT = PD - PT \tag{3}$$

Logo, as PNT em um sistema de distribuição são a diferença entre PD, um valor calculado a partir de medições, e PT, um valor calculado a partir das perdas técnicas em cada segmento. A partir dessa afirmação, é possível concluir que a precisão da estimativa das PNT é altamente dependente da qualidade dos modelos e precisão das variáveis utilizados no cálculo das PT, bem como na precisão das medições e estimativas de energia fornecida e faturada.

2.3 Perdas Técnicas

As PT são definidas como aquelas perdas inerentes ao transporte, transformação e medição de energia. Entre os efeitos causadores desse tipo de perdas pode-se citar (MÉFFE, 2001):

- efeito Joule nos condutores;
- correntes de fuga em isoladores;
- correntes de fuga em pára-raios;
- perdas em transformadores: perdas-cobre nos condutores (efeito Joule), perdas-ferro no núcleo (histerese, correntes induzidas) e indução de corrente na carcaça;
- efeito corona em linhas de alta tensão;
- medidores de energia;
- capacitores;
- reguladores de tensão;
- ramais de ligação;
- conexões;

Segundo o PRODIST, o cálculo das perdas é realizado por segmento e deve ser apresentado pelas distribuidoras (AGÊNCIA NACIONAL DE ENERGIA ELÉTRICA, 2013a). As PT são calculadas para os segmentos de rede, transformação, ramal de ligação e medidor por modelos aproximados. Além disso, a metodologia de cálculo é diferente para segmentos na alta, média e baixa tensão. Nos sistemas de alta tensão há abundância de medidores de grandezas elétricas, logo é possível obter as perdas de potência e energia diretamente a partir do sistema de medição. Para os demais níveis de tensão, as perdas são calculadas a partir de métodos específicos para cada segmento com base na potência média consumida e em parâmetros do sistema. Uma vez obtidas as PT em potência, elas são então convertidas para perdas em energia por meio do coeficiente de perdas.

Há uma série de simplificações adotadas na estimativa de PT. O PRODIST, na seção 7.2, define que as perdas de potência em um sistema de distribuição são calculadas com

base na demanda média, considerando o efeito das variações temporais da potência na estimativa por meio do coeficiente de perdas apenas(AGÊNCIA NACIONAL DE ENERGIA ELÉTRICA, 2013a). Diferentes modelos de carga também não são considerados e podem causar erros consideráveis no cálculo de PT (ROSSONI et al., 2013; DRESCH, 2014). O fator de potência considerado para todos os cálculos é de 0,92 e a tensão considerada é sempre a nominal. Considera-se que as cargas são equilibradas no sistemas de média tensão e desequilibradas para o sistema de baixa tensão. As perdas adicionais causadas pelo desequilíbrio de carga nas fases e pelo posicionamento assimétrico do transformador em relação às tipologias da rede são contabilizadas pelo acréscimo de 15% sobre as PT calculadas no sistema de baixa tensão. São considerados os níveis de tensão nominal de operação de cada distribuidora. Além disso, algumas fontes de PT são de difícil apuração, tais como aquelas produzidas por efeito corona, sistemas supervisórios, relés fotoelétricos, capacitores, transformadores de corrente e de potencial, e por fugas de correntes em isoladores e pára-raios. Estas são consideradas pelo acréscimo de 5% sobre o montante total de PT calculadas (AGÊNCIA NACIONAL DE ENERGIA ELÉTRICA (Brasil), 2013a).

Também é importante ressaltar que as PNT são causas de PT não contabilizadas. No calculo de PT leva-se em consideração apenas a potência média, calculada a partir da energia medida. Alguns métodos encontrados na literatura tentam refinar a estimativa de PT incluindo no seu cálculo a parcela de PNT. Como não é possível saber *a priori* a localização das PNT, uma estratégia adotada é distribuir as PNT proporcionalmente às cargas de maneira iterativa, até que a igualdade em (3) seja respeitada dentro de uma margem de erro aceitável ou de maneira direta, por uma estimativa baseada na diferença entre PD e PT (MÉFFE, 2007; MÉFFE; de Oliveira, 2009).

2.4 Perdas Comerciais

Diferentemente das PT, as PC são causadas por ações externas ao sistema de potência, não sendo, portanto, inerentes a este (SURIYAMONGKOL, 2002, p. 2). As principais causas de PC são (AGÊNCIA NACIONAL DE ENERGIA ELÉTRICA (Brasil), 2014; SURIYAMONGKOL, 2002; QUEIROGA, 2005):

- furto de energia elétrica;
- fraudes em medidores;
- auto-religações;
- defeitos em medidores;

O furto de energia elétrica é caracterizado pela utilização de derivações no circuito elétrico de maneira que a carga atendida por esses circuitos não seja medida. Esse tipo de irregularidade pode ser facilmente realizado em sistemas de baixa tensão e não exige a violação do medidor (SURIYAMONGKOL, 2002, p. 39). Em casos onde o sistema de distribuição é subterrâneo, a identificação deste tipo de ligação clandestina necessita de ferramentas especializadas.

As fraudes em medidores ocorrem quando há violação ou adulteração destes equipamentos com objetivo de reduzir a quantidade de energia medida (QUEIROGA, 2005). Em geral, esse tipo de irregularidade requer o rompimento dos lacres dos medidores para a adulteração de seus componentes internos. Uma das maneiras de se fraudar um medidor

eletromecânico é obstruir o disco ou o seu eixo a fim de parar a contagem de energia ou reduzir a velocidade do disco (SURIYAMONGKOL, 2002, p. 37). A adulteração dos enrolamentos dos elementos medidores de corrente ou o seu curto-circuito são uma maneira comum de fraudar o medidor. Além de violações no equipamento, adulterações na tensão de neutro, utilizando um transformador, por exemplo, podem reduzir a quantidade de energia medida (SURIYAMONGKOL, 2002, p. 39).

A auto religação ocorre quando o fornecimento de energia de um consumidor é suspenso, em geral por falta de pagamento, e antes da regularização da sua situação junto à distribuidora ocorre uma religação, portanto clandestina, do mesmo à rede elétrica. Quando há falha na medição cuja causa não é uma ação deliberada de um indivíduo, considera-se que há defeito no medidor. O não pagamento da fatura pelos consumidores é citado como uma das maiores causas de prejuízo financeiro (SURIYAMONGKOL, 2002, p. 5).

Outras causas de PC incluem (AGÊNCIA NACIONAL DE ENERGIA ELÉTRICA (Brasil), 2014; SURIYAMONGKOL, 2002; QUEIROGA, 2005):

- erros na medição inerentes à incerteza nos valores aferidos pelos medidores;
- erros na estimação do consumo de consumidores que não possuem medidores ou cujos medidores apresentaram defeito;
- erros no faturamento;
- ausência de medidores energia;
- erros no cálculo de PT;
- erros em cadastros de dados que causam imprecisões no cálculo de perdas técnicas;

2.5 Conclusão

Enquanto as PT são inerentes ao transporte de energia elétrica e as maneiras de calculalas são conhecidas, é virtualmente impossível medir diretamente as PC em um sistema de distribuição. É evidente que a apuração de PNT é bastante imprecisa por conta da grande quantidade de estimativas e simplificações adotadas para tornar possível a apuração de PT. A partir dessa constatação, pode-se afirmar que metodologias para apuração de PC baseadas em métodos como fluxo de carga certamente necessitam de fatores de correção para levar em conta a carga adicional gerada pelas PC e as PT por elas causadas. Mesmo com fatores de correção, as metodologias mencionadas neste capítulo não são capazes de localizar as PC, apenas de refinar a estimação do seu montante.

Dentre os tipos de PC estudados, alguns como aqueles causados por erros em parâmetros utilizados para cálculo de PT e erros em dados de cadastro podem ser mitigados pela utilização de equipamentos mais precisos e revisões nas bases de dados e procedimentos das empresas distribuidoras de energia elétrica. Outros deles só podem ser averiguados por inspeção, como é o caso das auto-religações e na ausência de medidor na UC. No entanto, os principais, que são os furtos de energia, as fraudes e os defeitos em medidores podem, em alguns casos, ser diagnosticados a partir da análise dos dados provenientes das leituras dos medidores. A partir da bibliografia consultada, pode-se concluir que um efeito direto da inserção dos principais tipos de PC é a redução no consumo mensal medido.

3 DETECÇÃO E IDENTIFICAÇÃO DE PERDAS COMERCIAIS

3.1 Introdução

Uma vez que as PC não são causadas diretamente pelo processo de transporte de energia elétrica, elas podem ser evitadas. Conforme (AGUERO, 2012), as principais técnicas de redução de PC estão relacionadas a impedir e identificar a ocorrência de furtos ou fraudes.

As técnicas que tem como base impedir a ocorrência de furtos e fraudes em geral tem como base a implementação de ações sobre os medidores e condutores. A mais elementar é a instalação de medidores em todos os consumidores e empreendimentos, independentemente do tipo de UC ou da modalidade da tarifa. Desta maneira, pode-se reduzir os erros nos cálculos de PC (AGUERO, 2012). Outra alternativa implementada com sucesso é a utilização de medidores pré-pagos. A instalação de medidores coletivos para o medição e cálculo de PT e PC em locais como o secundário de TD, auxilia na localização e contabilização precisa de PC e também mostrou bons resultados em um projeto-piloto, reduzindo o nível de PC de 25% para 17% (IGLESIAS, 2006). A utilização de medidores inteligentes, também chamados de *advanced metering infrastructure* (AMI), com funcionalidades tais como alarmes anti-fraudes. A implementação desses tipos de medidores também adiciona capacidade de embutir no próprio medidor algoritmos capazes de realizar análise de comportamentos anômalos de consumidores, apontando a ocorrência de fraudes (AGUERO, 2012). Entre as medidas que utilizam condutores, podemos citar a utilização de condutores de baixa tensão subterrâneos e condutores anti-fraudes que geram curto-circuitos quando cortados.

Já os métodos de identificação a ocorrência de PC se valem de informações para detectar padrões ou encontrar áreas onde há grandes quantidades de PC. Como é relatado na literatura, um dos métodos mais tradicionais de identificação de PNT é o degrau de consumo (PENIN, 2008). O degrau de consumo é proveniente da análise do histórico de consumo de uma UC. Quando o consumo de uma UC é reduzido significativamente por alguns meses consecutivos, está caracterizado o degrau de consumo. Esse é um dos métodos tradicionais de identificação de PC utilizados pelas distribuidoras e que estão baseados na perícia dos especialistas da empresa. A classificação de clientes suspeitos a partir da análise dos seus dados de fatura e cadastrais realizado por especialistas de empresas de distribuição é uma prática pouco eficiente. Na literatura, as experiências deste tipo documentadas apontam que as suspeitas apontadas por esse tipo de método eram apenas confirmadas em 13% a 15% (HERNANDES JR et al., 2013, p. 48), 25% (ORTEGA, 2008, p. 24) e de 7% a 25% (QUEIROGA, 2005, p. 74) dos casos.

A inspeção em campo dos consumidores por equipes técnicas é a forma mais eficaz de se identificar as PC. Porém, não é financeiramente viável enviar equipes para inspecionar todos os consumidores. Na literatura especializada são encontrados em grande número métodos visando realizar a detecção e identificação de PC em sistemas de distribuição. O objetivo desses métodos é localizar UC onde haja fraudes no sistema de medição, furtos de eletricidade na rede elétrica, medidores de energia defeituosos ou ausentes que representam prejuízo econômico às empresas distribuidoras de energia sob a forma de energia não faturada e acréscimo de perdas técnicas. Dentre esses métodos, podemos destacar a utilização de estimadores de estado para a localização de áreas onde há maior quantidade de PC (CRUZ; QUINTERO; PÉREZ, 2006). Outro método utilizado com sucesso é baseado na utilização de medidores portáteis em miniatura para procedimento de pré-inspeção. Esses medidores são instalados de forma discreta no ramal de ligação de UC e registram o seu consumo durante algum tempo. Depois a leitura no medidor instalado na UC é confrontada com aquela do medidor portátil e se há grandes discrepâncias entre as duas medições, uma inspeção é realizada. Em projetos-piloto, a taxa de acerto no reconhecimento de UC irregulares e regulares foi de 100% (HERNANDES JR et al., 2013).

Uma alternativa que tem sido amplamente utilizada é o uso de métodos baseados em inteligência artificial com o intuito de identificar padrões de consumidores suspeitos a partir de bases de dados contendo informações sobre as UC (RAMOS et al., 2011). Alguns métodos são capazes de apontar UC fraudadoras, enquanto outros indicam a probabilidade de se encontrar um fraudador em um agrupamento de UC. A vantagem destes métodos é a redução do universo de UC candidatas a inspeção a uma lista na qual a probabilidade de se encontrar PC em uma inspeção é elevada. Em posse dessa informação, as distribuidoras podem, de maneira eficiente, realizar inspeções em clientes para detecção de PNT, reduzindo assim os prejuízos por fraudes e furto de energia elétrica. O envio de equipes de inspeção a campo é uma atividade custosa à empresa e que em geral possui baixo índice de sucesso, daí o interesse em obter-se um bom método de inferência de perfis fraudadores.

Este trabalho propõe a utilização de métodos baseados em inteligência artificial para localização e identificação de PC. Neste capítulo serão apresentados em mais detalhe os principais métodos encontrados na literatura atualmente (seção 3.2). São eles: Redes Neurais Artificiais (subseção 3.2.2), Máquinas de Vetor de Suporte (subseção 3.2.3) e Floresta de Caminhos Ótimos (subseção 3.2.4). Na seção 3.3 são detalhados quais dados podem ser utilizados no projeto de classificadores deste tipo de acordo com as normativas da ANEEL. Além disso, são apresentadas as métricas para avaliação de desempenho utilizados neste trabalho (seção 3.4).

3.2 Inteligência Artificial

Entre os métodos empregados para detecção de PC encontrados na literatura, aqueles que utilizam inteligência artificial para apontar possíveis fraudes ou defeitos em medidores em sistemas de distribuição são os mais numerosos. Esses trabalhos tem demonstrado que, com o auxílio de programas de computador dotados de capacidade de aprendizado, é possível aumentar a chance de encontrar UC onde há PNT por meio do reconhecimento de padrões de fraude ou defeito em medidores. Sistemas utilizando inteligência artificial podem ser capazes de detectar automaticamente novos padrões bem como buscar padrões já conhecidos por experiência humana.

Algumas técnicas empregadas são: regressão logística (PENIN, 2008), análise linear

discriminante (ou *Linear Discriminant Analysis* – LDA) (PENIN, 2008), máquinas de vetor de suporte (support vector machines, SVMs) (DEPURU; WANG; DEVABHAKTUNI, 2011; NAGI et al., 2008a, 2011, 2010; NIZAR; DONG, 2009), redes neurais artificiais (*Artificial Neural Networks*– ANN) (NIZAR; DONG, 2009; DEPURU et al., 2011; ORTEGA, 2008; PENIN, 2008; QUEIROGA, 2005) , redes bayesianas (QUEIROGA, 2005; BASTOS; SOUZA; FERREIRA, 2009a,b; MONEDERO et al., 2012), árvores de decisão (QUEIROGA, 2005; MONEDERO et al., 2012), *Extreme Learning Machines* (NIZAR; DONG; WANG, 2008), *clustering* ou análise de conglomerados (ANGELOS et al., 2011; NIZAR; DONG, 2009) e floresta de caminhos ótimos (*Optimum-Path Forest* – OPF) (RAMOS et al., 2011, 2009, 2012, 2011). O artigo (NAGI et al., 2010) propõe um método que aumentaria a taxa de sucesso em inspeções de 3% para 60% em uma distribuidora de eletricidade da Malásia após a adoção de um sistema inteligente de detecção de perdas não técnicas baseado em SVMs. Em um artigo mais recente, a taxa se sucesso sugerida chegou a 72% após aperfeiçoamento do sistema (NAGI et al., 2011).

3.2.1 Sistemas de Reconhecimento de Padrões

O problema da detecção e identificação de perdas não-técnicas em sistemas de distribuição foi abordado na literatura por muitos autores como um problema de reconhecimento de padrões. Segundo (DUDA; HART; STORK, 2000), o reconhecimento de padrões pode ser definido como o processo de receber dados não tratados e tomar uma ação com base no padrão reconhecido.

No caso dos sistemas que se propõem a detectar PNT, a entrada são dados de unidades consumidoras UC ou grupos de UC, como consumo mensal, localização geográfica e resultado de inspeções. A parte de percepção é a responsável por adquirir esses dados e formar uma base de dados, como o cadastro de UCs e medição de grandezas elétricas no caso de distribuidoras de energia elétrica. Na literatura há diversos exemplos de dados utilizados. Em (PENIN, 2008), a base de dados utilizada possui código do cliente, dados de resultados de inspeção, data da detecção de irregularidade, consumo base da UC, classe de atividade identificada pela equipe de inspeção de campo, atividade identificada e consumos mensais indicados. Já em (ANGELOS et al., 2011) foi proposto um sistema de detecção de PNT baseado no consumo de UC, em observações de inspeções e na localização da UC. Em (ORTEGA, 2008), uma extensa base de dados foi utilizada. Os dados considerados incluíam: local da inspeção, origem da inspeção, consumo do cliente inspecionado, cortes do cliente inspecionado, equipamento de medição, inspeções, temperatura, curva de consumo, dados cadastrais de cada cliente, classe de consumo, carga instalada, para citar alguns.

A segmentação pode ser vista, nesse caso, como a eliminação dos dados de UC que não são de interesse. A extração de características é a transformação dos dados selecionados em valores numéricos (as medidas citadas anteriormente, geralmente são números reais ou inteiros) formando vetores para cada amostra (neste caso são as UC). Esses vetores terão um sentido matemático, estando inseridos em um espaço espectral de dimensão igual ao número de características selecionadas. Na Figura 2 temos um exemplo de um problema binário e as amostras são caracterizadas por um vetor bidimensional da forma $v = [f_1 f_2]^T$. No caso da detecção de PNT, o problema é igualmente binário, pois existem as classes regular e irregular, mas a dimensão do espaço espectral é maior. Em (ANGELOS et al., 2011), são utilizadas 5 dimensões, com indicadores como médias móveis de consumo mensal, picos de consumo mensal, desvio padrão, soma do número de observações em inspeções e dados de consumo médio da região em que a UC é localizada. Já em

(ORTEGA, 2008), a dimensão do espaço espectral é muito maior, com 26 atributos.

A parte de classificação é aquela que é geralmente citada nos artigos relacionados a detecção de PNT, pois geralmente os classificadores utilizados são ferramentas já bastante estabelecidas e em geral não há contribuições em termos de novos classificadores nos estudos apresentados. Diferentemente das partes citadas anteriormente, essa é aquela cujo projeto é mais discutido na literatura, e é onde a parte de inteligência artificial fica mais evidente.

Figura 1: Etapas de um sistema de reconhecimento de padrões. Fonte: (DUDA; HART; STORK, 2000).

O projeto do classificador é chamado de treinamento. Existem 3 maneiras de se treinar classificadores: aprendizado supervisionado, aprendizado não-supervisionado e aprendizado por reforço (DUDA; HART; STORK, 2000). Nem todos os classificadores apresentados anteriormente possuem a capacidade de aprender das três maneiras distintas, em geral apenas uma delas. Uma vez treinado, supõe-se que o classificador tem a capacidade de prever a qual classe uma amostra pertence, apenas utilizando como critério as características extraídas para cada amostra. Resumidamente, o aprendizado supervisionado é similar ao aprendizado com um professor. A partir de um grupo de amostras cuja classe é conhecida previamente, projeta-se um classificador capaz de separar o espaço amostral em regiões diferentes para cada classe com erro mínimo de classificação para o grupo de treinamento dado. Então, supõe-se que aplicando o mesmo critério para um grupo de amostras cuja classe é desconhecida a princípio, o resultado será parecido com o resultado da classificação para o grupo de treinamento. Logo, para que o aprendizado supervisionado seja possível, é necessário conhecer certo número de amostras antes de realizar o projeto do classificador. Esse é o caso para muitos trabalhos na literatura (DEPURU; WANG; DEVABHAKTUNI, 2011; NAGI et al., 2008a, 2010; DEPURU et al., 2011; ORTEGA, 2008; PENIN, 2008; QUEIROGA, 2005; BASTOS; SOUZA; FERREIRA, 2009a,b; MO-

NEDERO et al., 2012; NIZAR; DONG; WANG, 2008; RAMOS et al., 2011, 2009, 2011; LEÓN et al., 2011; NIZAR et al., 2007). Para a avaliação do desempenho de classificador, utiliza-se um grupo de teste, que possui amostras cujas classes são conhecidas, mas que não foram utilizadas no projeto de classificador. As amostras desse grupo de teste são utilizadas como entrada para o classificador e avalia-se então a precisão do classificador, ou seja, qual foi o acerto nas classificações.

No caso do aprendizado não supervisionado, não há necessidade de um grupo de teste com classes conhecidas previamente, mas de um critério para o agrupamento (*clustering*) das amostras do grupo de treinamento. Esse critério pode ser, por exemplo, distância euclidiana entre os pontos que representam cada amostra no espaço amostral. Diferentes critérios resultam em diferentes *clusters*. Isso dá origem a classes obtidas de maneira "natural"(DUDA; HART; STORK, 2000). Em alguns artigos, foi utilizado o critério de distâncias euclidianas para se formarem *clusters* que foram classificados como UC com comportamento regular, e os pontos aberrantes (*outliers*) foram considerados como UC candidatas a inspeção (ANGELOS et al., 2011). Métodos que utilizam aprendizado por reforço, também conhecido como aprendizado com um crítico, não foram encontrados na literatura relacionada à detecção de PNT. Trata-se de um tipo de aprendizado em que, dadas as características da amostra a ser rotulada, o classificador prevê uma classe de acordo com critérios estabelecidos previamente. É então avaliado o acerto ou erro do classificador e essa informação realimenta os critérios de classificação para melhorar o seu desempenho (DUDA; HART; STORK, 2000). A sua desvantagem com relação ao aprendizado supervisionado é que os padrões não são aprendidos diretamente por meio das amostras rotuladas mas são determinados a partir de sucessivas "críticas".

Ainda de acordo com (DUDA; HART; STORK, 2000), o projeto de sistemas de reconhecimento de padrões é um processo iterativo, composto pelas tarefas de aquisição de dados, escolha de características, escolha de modelo, treinamento e avaliação.

Florestas de Caminhos Ótimos (OPF) e Máquinas de Vetor de Suporte (SVM) são dois classificadores com conhecida utilização para o problema de identificação de perdas comerciais em sistemas elétricos de potência (RAMOS et al., 2009) (RAMOS et al., 2011) (RAMOS et al., 2012) (NAGI et al., 2011) (DEPURU; WANG; DEVABHAKTUNI, 2011) (NAGI et al., 2010). Ambos possuem versões supervisionadas e não supervisionadas. Por conta dos resultados bastante satisfatórios apresentados nos artigos, escolheu-se dar ênfase a esses dois nas pesquisas. No caso das SVM utiliza-se um hiperplano para separar duas classes. Caso elas não sejam separáveis, transformações não lineares são realizadas de maneira a aumentar a dimensão do espaço amostral considerado até tornar as classes separáveis.

Já no caso do OPF, as amostras são modeladas como nós de um grafo e o conhecimento necessário para o bom entendimento do método não depende apenas de conhecimentos de geometria, mas principalmente de conceitos da teoria de grafos. Inicialmente considera-se um grafo completo, que ao fim da fase de treinamento se torna uma floresta de caminhos ótimos, de acordo com critério de mínimo custo (utilizando a função f_{min}) (PAPA et al., 2007). Cada caminho dessa floresta pertencerá a uma classe. As amostras a ser classificadas são inseridas uma a uma nessa floresta para classificação e depois de classificadas são removidas da floresta. Ou seja, o classificador não se altera durante o processo de classificação. Uma vez inserida no grafo, a amostra é conectada por arestas a todos os nós da florestas, cada uma com um custo. O caminho que apresentar menor custo de acordo com a função f_{min} passa seu rótulo à amostra testada, realizando assim a sua classificação.

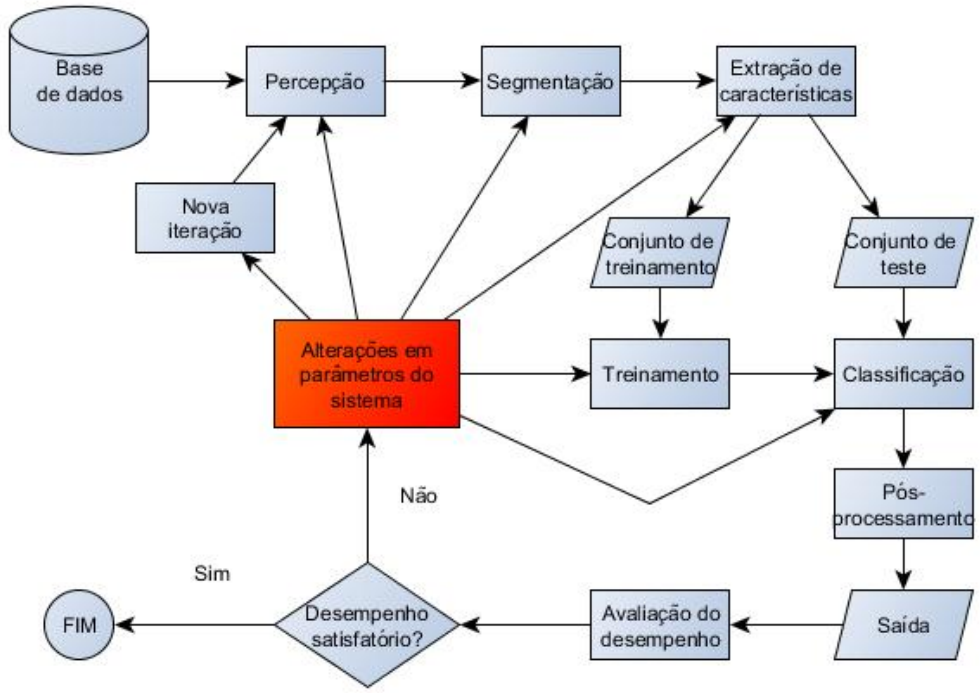

Figura 2: Etapas do projeto sistema de reconhecimento de padrões. Fonte: (DUDA; HART; STORK, 2000) (com adaptações).

No pós-processamento os resultados da classificação são utilizados para se tomar decisões e são tratados com o objetivo de cumprir os objetivos do sistema de reconhecimento de padrões. Nessa parte incorporam-se outros critérios ao projeto de classificador (portanto pode haver realimentação no projeto do classificador nesta parte) como a substituição do critério de erro mínimo do classificador por critérios de custo ou risco, mais adaptados à aplicação (DUDA; HART; STORK, 2000). Em sistemas com vários classificadores, nesta etapa reúne-se os resultados da classificação e decide-se por uma ação. Em (ORTEGA, 2008), utilizou-se um critério de comitês para se criar a lista de UCs candidatas a inspeção, a partir dos resultados de diferentes redes neurais treinadas com grupos de treinamento distintos.

3.2.2 Redes Neurais Artificiais

São algoritmos que buscam realizar o processamento de dados mimetizando uma rede neural natural, como por exemplo, o cérebro humano. O sistema é formado pela conexão entre neurônios artificiais, que são a unidade básica de processamento de dados.

Cada neurônio é capaz de realizar operações matemáticas simples a partir de várias entradas e com apenas uma saída, de acordo com uma função de ativação. Essa função é normalmente não linear, sendo ativada quando a soma ponderada das entradas do neurônio, chamada de nível de atividade, excede certo patamar (ORTEGA, 2008).

As RNA são caracterizadas pelo padrão de conexão entre os neurônios, pelo método de determinação dos pesos das conexões (treinamento) e pela função de ativação. Podem existir redes de camada simples, nas quais cada saída é resultado do processamento dos dados por um neurônio, e redes multicamadas, assim classificadas quando os neurônios que geram as saídas da RNA processam informações geradas por outros neurônios. Podem ainda existir redes neurais realimentadas (ORTEGA, 2008).

O treinamento estático das RNA é basicamente um processo onde os pesos das conexões (ponderação das entradas) são ajustados até que a RNA seja capaz de fornecer uma solução generalizada para certa classe de problemas. Já o treinamento dinâmico, pode ocorrer mudança no número de conexões, camadas e neurônios (QUEIROGA, 2005). O processo de treinamento pode ocorrer de forma supervisionada ou não supervisionada. O aprendizado supervisionado é dado quando há um professor que conhece o comportamento do ambiente e responde aos estímulos deste da forma desejada. Assim, uma função de aprendizado age de forma iterativa, ajustando os pesos das conexões até que a rede neural seja capaz de responder ao estímulo do ambiente de forma mais parecida quanto for possível com a forma que o professor reage. No caso do aprendizado não supervisionado, não há professor e os neurônios agem como classificadores e as suas entradas são os elementos a serem classificados. Somente é ativado aquele neurônio cujos pesos são mais próximos ao valor de entrada. Assim, através de um processo de competição e cooperação entre os neurônios, cada neurônio é treinado por uma regra de aprendizado de maneira a ser capaz de classificar uma classe de padrões do sinal de entrada (ORTEGA, 2008).

Como principais características positivas pode-se citar a adaptabilidade, o processamento de dados não lineares e sujeitos a ruído e o processamento paralelo. As desvantagens são que as RNA requerem treinamento, os dados de entrada são difíceis de serem formatados e os modelos produzidos são de difícil entendimento. No trabalho apresentado por (PENIN, 2008), as redes utilizadas foram as Redes Neurais Perceptron Multicamadas, e as Redes de Kohonen. Em (ORTEGA, 2008) foram utilizadas Redes Neurais Perceptron Multicamadas e também uma técnica chamada de comissionamento, que consistem em criar múltiplas redes neurais que agem conjuntamente na forma de comitês, com o objetivo de melhorar a generalização do sistema.

Em (ORTEGA, 2008) foi realizado um pré-processamento da extensa base de dados da companhia Light, do qual foram extraídas para cada UC 32 variáveis usadas como entrada da RNA, entre indicadores, classificadores e dados brutos. O índice de acerto dos clientes apontados como suspeitos variou entre 24,9% a 63,9% nos casos estudados, ao aplicar-se o Módulo de Classificação descrito no trabalho. Em (QUEIROGA, 2005), o algoritmo de RNA utilizado foi um dos que obteve melhores resultados na identificação de clientes fraudadores, alcançando taxas de acerto na classificação de clientes irregulares de 59% no melhor dos casos.

Um aspecto importante sobre a aplicação de redes neurais na identificação de perdas comerciais é que não há necessidade de cálculo de perdas técnicas para aplicação dessas técnicas. Por outro lado, outros dados como histórico e dados de cadastro de cliente são necessários. Em (ORTEGA, 2008) e em (QUEIROGA, 2005) base de dados utilizada é muito mais extensa que em (PENIN, 2008). Neste, os resultados foram inconclusivos enquanto nos outros dois ficou comprovado que o aprendizado da RNA implementada melhora significativamente o a taxa de acerto na classificação de clientes apontados como suspeitos para todos os casos mostrados. Um dos limitantes dessa técnica é que a identificação dos clientes irregulares é altamente dependente dos dados utilizados no treinamento, que são obtidos através das inspeções das UC, o que caracteriza aprendizado supervisionado. Logo, as redes neurais aprendem a identificar apenas os padrões dos clientes irregulares dentro do grupo de clientes julgados como possíveis fraudadores e então submetidos a inspeção. Pode-se criticar essa metodologia por limitar o aprendizado de reconhecimento de padrões a uma base de treinamento que não forma uma amostra estatística significativa. Uma alternativa seria obter dados de inspeções realizadas de forma aleatória para que se obtenha uma medida estatística mais significativa ou empregar mé-

todos de aprendizado não supervisionado.

3.2.3 Máquinas de Vetor de Suporte

As *Support Vector Machines* (SVM), também conhecidas como Máquinas de Vetores de Suporte, são uma técnica de aprendizado de máquina que vem sendo utilizada com sucesso em detecção e localização de PNT principalmente por ter uma capacidade de fornecer soluções de grande generalidade para problemas de classificação de padrões (NAGI et al., 2011) (DEPURU; WANG; DEVABHAKTUNI, 2011) (NAGI et al., 2010). Também pode ser utilizado para realizar regressão não-linear (HAYKIN, 1998). A classificação é uma função não-linear, cuja entrada são as características das amostras e a saída é a classe à qual pertencem. A formulação original das SVM trata de problemas de classificação binária, mas pode ser estendida para problemas de classificação com múltiplas classes a um custo computacional elevado.

As SVM podem realizar aprendizado supervisionado ou não supervisionado. No primeiro caso, elas realizam agrupamento (*clustering*) de dados de acordo com certo critério de qualidade. No segundo caso, que é o caso de interesse, elas aprendem a realizar a classificação a partir de um conjunto de treinamento, que são as amostras analisadas. Os dados de entrada, no caso de identificação e localização de PNTs podem ser dados cadastrais, histórico de consumo ou mesmo informações obtidas a partir de curvas de carga. A saída é a classificação em UC com ou sem perdas comerciais.

Esse método é uma implementação aproximada do método de minimização de risco estrutural, princípio fundamentado na Teoria de Aprendizado Estatístico (HAYKIN, 1998). Essa teoria de aprendizado trata do problema fundamental de como controlar a capacidade de generalização de uma máquina de aprendizagem em termos matemáticos (HAYKIN, 1998). Seguindo o princípio do menor risco estrutural, que diz que o erro de um classificador é função da complexidade das suas funções de classificação (quantificada pela dimensão Vapnik-Chervonenkis), as SVM utiliza um classificador linear, que em geral é um hiperplano que realiza a classificação binária dividindo em dois os conjuntos de dados. Esses hiperplanos são otimizados de maneira que se obtenha aquele que possui a maior distância (margem de separação) entre os pontos de classes diferentes (HAYKIN, 1998). Esses pontos que estão à mínima distância do hiperplano de classificação são os chamados vetores de suporte, que forma um pequeno subconjunto dos dados de treinamento chamado.

Para tratar de *outliers*(valores atípicos ou aberrantes) e funções não linearmente separáveis, é preciso tornar a SVM tolerante a erros de classificação. Assim, o problema da definição do hiperplano ótimo é redefinido não mais para maximizar a distância entre os pontos de classes diferentes, mas para minimizar a probabilidade de erro de classificação. No entanto, geralmente, os problemas não são do tipo linear, ou seja, não podem ser corretamente classificados unicamente através de um hiperplano. Para resolver esse problema, utilizam-se transformações, em geral não lineares, na entrada que aumentam a sua dimensão de maneira que o problema possa admitir classificação por hiperplano. A função que realiza esta transformação é chamada *kernel*. Em muitos casos pode ocorrer um grande aumento na dimensão do problema, causando grande custo computacional. Além disso, em alguns casos não é possível separar dados apenas para um número finito de dimensões, e nesses casos o método é ineficaz (RAMOS et al., 2009).

Um dos problemas relacionados com as SVM é que elas não fornecem ao usuário regras de classificação, como acontece com as árvores de decisão ou com métodos baseados em *clustering*, por exemplo. Na referência (DEPURU; WANG; DEVABHAKTUNI,

2011) foi citada a utilização de uma biblioteca em MATLAB que implementa SVM, chamada SVMLIB (CHANG; LIN, 2001). Esta biblioteca utiliza como *kernel* a função de base radial gaussiana:

$$K(x_i, x_j) = e^{-\gamma \|x_i - x_j\|} \tag{4}$$

As UC foram separadas de acordo com as seguintes categorias: localização geográfica, estação do ano, classificação de consumidores por tamanho e setor. Os dados analisados foram consumo de energia e perfis médios de carga. Neste artigo a métrica para avaliação do desempenho do classificador é o total de amostras classificadas corretamente, alcançando um índice de 98,4% para o método adotado.

Em (NAGI et al., 2010), também é utilizado o perfil de energia mensal como indicativo da existência de fraudes em UC e a mesma biblioteca SVMLIB é utilizada. A base de dados contém informações de consumo mensal em kWh, data da leitura do medidor, tipo de leitura do medidor, informação sobre furtos de energia, sobre inspeções em campo, localização geográfica, entre outros indicadores da empresa distribuidora, Tenaga Nasional Berhad. Neste trabalho, a taxa de acerto na classificação dos clientes classificados como suspeitos é de 77,41% para o método de SVM. No entanto, em testes realizados em campo com a lista de UC suspeitas de fraude obtidas pelo método, obteve-se apenas 26% de acerto, que é uma evolução com relação à taxa de acerto média das inspeções realizadas pela empresa, de 3%. Uma etapa de pós-processamento de dados utilizando um sistema de decisão aumentou essa taxa de 26% para 64%. Esse sistema de decisão é baseado no conhecimento humano sobre UC irregulares. Em um trabalho posterior (NAGI et al., 2011), esse índice pôde ser aumentado de 60% para 72% ao se utilizar um sistema de inferência *fuzzy* como método de pós-processamento.

3.2.4 Classificador OPF

É um *framework* para desenvolvimento de classificadores criado em 2005. Esse classificador foi utilizado para identificação de UC com irregularidades em (RAMOS et al., 2009), (RAMOS et al., 2011) e (RAMOS et al., 2012). Entre as vantagens deste método destacam-se etapa de treinamento com baixo esforço computacional (comparando com RNA ou SVM), ausência de parâmetros e realização classificação de forma implícita (RAMOS et al., 2009).

O *Optimum-Path Forest* (OPF) é baseado na teoria de grafos e sua aplicação inicial foi o reconhecimento de padrões em imagens, modelado como um problema de partição de um grafo. O objetivo é fazer a segmentação do grafo em dois, de modo que as amostras de cada classe de dados fiquem em suas respectivas partições. O OPF possui três versões: a versão com treinamento não supervisionado, a versão com treinamento supervisionado usando grafo k-NN e a versão com treinamento supervisionado usando grafo completo. Um grafo k-NN é aquele em que cada nó está conectado por uma aresta apenas aos k nós de menor distância euclidiana, onde k é um número inteiro arbitrário. Já um grafo completo é aquele em que todos os nós estão conectados uns aos outros por uma aresta, portanto neste caso todos os nós do grafo são adjacentes. Este último tipo de grafo é aquele utilizado na versão original do OPF, a qual é considerada a que tem melhor desempenho, portanto é a que será abordada neste trabalho.

Grafos são definidos por dois conjuntos: o conjunto dos nós e o conjunto de arestas. Os vetores de características do espaço de entradas são modelados como nós do grafo. As arestas são definidas pelas relações de adjacência, que no caso de um grafo completo, conecta todos os nós entre si.

A segmentação é a fase de treinamento supervisionado do OPF e ocorre da seguinte maneira: são escolhidos nós-semente, chamados de protótipos, que competem entre si para conquistar amostras com o intuito de formar agrupamentos. Os protótipos tem função semelhante aos vetores de suporte presentes na SVM, pois são os elementos que melhor representam a sua classe. Eles são escolhidos heuristicamente e a sua escolha afeta fortemente o resultado do OPF. Um método proposto é o uso do algoritmo DMST (*Directed Minimum Spanning Tree*), que gera um subgrafo que é uma árvore que conecta todos os nós com n–1 arcos, onde n é o número de nós. Esse subgrafo tem a propriedade de ser a árvore cuja soma de todas as distâncias das arestas é mínima. Aqueles nós que forem conectados por arestas a nós da classe oposta à sua são aqueles mais próximos à região de fronteira entre as classes e são, portanto, os nós mais significativos, que serão escolhidos como protótipos.

É necessário que haja pelo menos um protótipo por classe. A conquista de amostras é o ato de projetar a classe de um protótipo em uma amostra ainda não classificada. A métrica para decisão dessas conquistas é uma função de custo de caminho suave que tem como entradas os pesos das arestas que ligam os protótipos a cada nó do grafo (PAPA; FALCÃO; SUZUKI, 2009) relacionada com os pesos dados pela distância de cada aresta, que pode ser uma função de distância euclidiana, por exemplo. Após todos os nós serem conquistados, cada protótipo dá origem a uma árvore de caminho ótimo. A presença de várias árvores, por conta da existência de vários protótipos, dá ideia de uma floresta, daí o nome do método.

Na fase de testes ou na fase de classificação, uma amostra por vez é inserida no grafo junto com a floresta e é verificado então qual das árvores a conquistou, classificando a amostra (PAPA; FALCÃO; SUZUKI, 2009). A amostra então é retirada do grafo e a classificação da amostra seguinte pode ser realizada.

No artigo (RAMOS et al., 2009) a técnica de OPF foi comparada com 3 outras: SVM usando kernel do tipo *Radial Basis Function* (SVM-RBF), SVM utilizando *kernel* linear (SVM-LINEAR) e RNA multicamada perceptron (ANN-MLP). Foram usados dados de uma distribuidora de energia elétrica brasileira para consumidores de médio a grande porte. Os dados avaliados foram: demanda contratada, demanda medida ou demanda máxima, fator de carga e potência instalada. Esses indicadores foram obtidos a partir de perfis de consumo com medidas de potência média obtidas com intervalos de 15 minutos. Mostrou-se que o desempenho dos métodos OPF e SVM-RBF foi semelhante e muito superior ao dos outros dois em termos de acerto na classificação. No entanto, o algoritmo OPF foi muito mais rápido na etapa de treinamento. Em uma simulação o TCC de OPF foi de 90,21% enquanto para SVM-RBF esse índice atingiu 88,93%, Um ponto importante sobre esse artigo é que não há menção a indicadores como VPP ou SENS nos resultados obtidos. Logo, apesar do alto índice de classificação, não é possível saber com certeza se o método consegue obter sucesso na identificação de UC com perdas não técnicas, embora os autores afirmem a superioridade de OPF frente aos outros algoritmos de estado-da-arte de IA.

Um aprofundamento do trabalho mostrado em (RAMOS et al., 2009) foi apresentado em (RAMOS et al., 2011), mostrando novas versões de algoritmos de aprendizado para tolerar erros de classificação, realizar a poda de ramos das árvores e aprendizado com grupo de teste. As conclusões e resultados são semelhantes ao do artigo supracitado, mas desta vez com vantagem maior do OPF sobre demais métodos de classificação.

3.3 Dados Utilizados

Segundo a Resolução Normativa 414 da ANEEL em seu artigo 145 (AGÊNCIA NACIONAL DE ENERGIA ELÉTRICA (Brasil), 2010), as distribuidoras devem manter uma base de dados contendo informações que devem ser atualizados periodicamente. Muitas destes dados podem ser considerados relevantes para ao problema de detecção e identificação de PNT, tais como:

- número da UC;
- endereço da UC;
- classe e subclasse da UC;
- data da primeira ligação da UC e do início do fornecimento;
- data do encerramento da relação contratual;
- tensão contratada;
- potência disponibilizada;
- carga instalada declarada ou prevista no projeto de instalações elétricas;
- valores de demanda de potência e de energia elétrica ativa, expressos em contrato, quando for o caso;
- informações relativas aos sistemas de medição de demandas de potência e de consumos de energia elétrica ativa e reativa, de fator de potência e, na falta destas medições, o critério de faturamento;
- históricos de leitura e de faturamento;
- registros das solicitações de informação, serviços, sugestões, reclamações e denúncias contendo o horário e data da solicitação e das providências adotadas;
- registros dos créditos efetuados na fatura em função de eventual violação dos indicadores e prazos estabelecidos;
- código referente à tarifa aplicável;
- informações referentes as inspeções/intervenções da distribuidora nos equipamentos de medição, violação de selos e lacres instalados nos medidores, caixas e cubículos;
- informações referentes a cobranças resultantes de deficiência na medição ou de procedimento irregular;
- contratos firmados com consumidor cuja unidade consumidora pertença ao grupo A.

Além disso, a distribuidora deve disponibilizar, para consulta em tempo real, históricos de leitura e de faturamento relativos aos últimos 13 ciclos de faturamento. As informações contidas no cadastro devem ser armazenadas pelo prazo mínimo de 60 ciclos consecutivos e completos de faturamento.

Conclui-se então que uma parte significativa dos dados necessários para a implementação do método já é previsto em regulamentações da ANEEL, o que possibilita o aproveitamento de dados já coletados sem necessidade de construir integralmente a base de dados e os instrumentos de medição. Além disso, tendo em vista que todas as distribuidoras no Brasil devem seguir as mesmas regras, pode-se pensar na aplicação do método de identificação de perdas não técnicas para qualquer distribuidora do país.

3.3.1 Normalização de Dados

Transformações de dados são procedimentos nos quais operações matemáticas são aplicadas aos valores das variáveis com o intuito de obter dados em uma forma mais adequada ao modelo que se está utilizando (FERREIRA, 2005). Entre as transformações mais comuns está a normalização de dados, que tem como objetivo homogeneizar a variabilidade das variáveis de uma base de dados, criando um intervalo de amplitude similar onde todas as variáveis irão residir (FERREIRA, 2005). Normalizar dados requer a projeção de suas variáveis de uma faixa para outra, o que introduz distorções e tendências. Essas tendências e distorções podem ser utilizadas para expor melhor o conteúdo da informação, mas podem também surtir o efeito contrário dependendo da natureza dos dados (PYLE, 1999). Entre os tipos de normalização mais comuns estão a normalização pelo desvio-padrão, apresentada em (5)e a normalização pela faixa de variação, tal qual descrito por (6). No primeiro, os dados transformados ficam semelhantes à forma de uma distribuição normal padronizada: a média dos dados transformados é zero e os seus valores são dados pelo desvio-padrão.

$$y = \frac{x - \mu}{\sigma} \tag{5}$$

Em que x é um vetor de características a ser normalizado, μ é a média os vetores de características do conjunto analisado, σ é uma matriz diagonal contendo o desvio-padrão para cada característica do mesmo conjunto de dados e y é o vetor de características normalizado. Já no caso da normalização pela faixa de variação, deseja-se representar cada variável no intervalo entre 0 e 1. Isso pode ser realizado pela transformação descrita pela equação (6).

$$y = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{6}$$

Em que x_{min} é um vetor contendo os menores valores de cada características e x_{max} é um vetor contendo os maiores valores de cada características. Como no caso do OPF a dissimilaridade entre duas amostras é dada pela função custo f_{max}, que também é função da distância euclidiana do caminho entre as amostras e um protótipo, a normalização certamente insere distorções na maneira de tratar os dados. Na literatura, ainda são descritos outros tipos de transformações, como o uso da função inversa $(1/x)$, logaritmo ou a extração de diferenças temporais (FERREIRA, 2005).

Tendo em vista as particularidades do problema abordado, percebeu-se que estas duas transformações não cumprem o objetivo de deixar evidente o degrau de consumo causado pelas PC, uma das suas características mais evidentes. Isso poderia ser implementado pela normalização pelo consumo médio, como em (7).

$$y = \frac{x}{x_{med}} \tag{7}$$

Em que x_{med} é a média das características do vetor x. Como pode ser observado na figura 3, a partir desta normalização é possível comparar os históricos e observar características semelhantes entre eles.

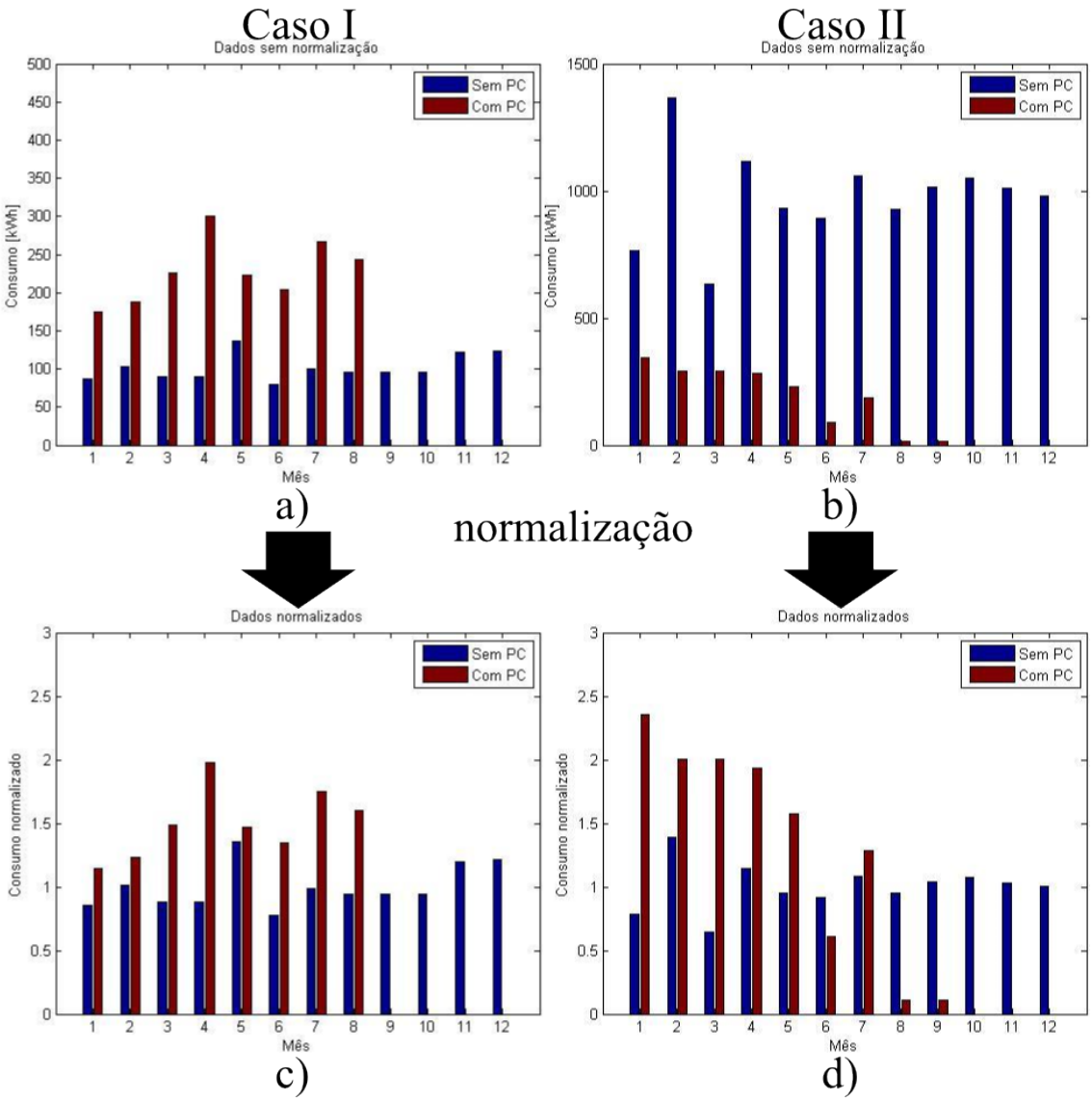

Figura 3: Efeitos da normalização de históricos mensais de consumo. Fonte: Próprio autor.

Nesta figura, são observados dois casos de comparação entre padrões de consumo mensal em UC onde não há PC, em azul, e outras dois em que há PC, em vermelho. Em (a) e (b) não há normalização pela média, enquanto em (c) e (d) são mostrados os dados normalizados. No Caso I, os consumos mensais médios dos dois clientes são próximos. Antes da normalização, em (a), a dissimilaridade entre os dois gráficos é maior, embora a tendência de demanda maior da curva vermelha nos primeiros 8 meses e da curva azul nos 4 últimos se mantenha em (c), havendo uma diferença clara entre elas. Já no Caso II, quando não há normalização, em (b), fica difícil comparar as duas curvas pois o consumo mensal médio delas é muito diferente. No entanto, quando há a normalização pela média, como em (d), elas se tornam comparáveis e o mesmo padrão do degrau de consumo fica

evidente, assim como em (a) e (c). Depois da normalização, observa-se que, por conta das PC, os históricos de consumo com PC tendem a possuir valores de consumo mensais bastante diferentes de 1 tanto para mais quanto para menos. Logo, pode-se concluir que a normalização ajuda na criação de superfícies de separação entre os elementos de classes distintas.

3.4 Avaliação de Desempenho

Para permitir a comparação entre diferentes métodos é necessário padronizar a apresentação dos resultados de cada um deles. A maioria dos métodos de identificação e localização de PNT parte de um conjunto de dados obtido a partir de inspeções realizadas em UC. Neste conjunto de dados, ocorre a classificação de cada UC ou cada conjunto de UC de acordo com o resultado das inspeções realizadas em campo.

Os resultados das inspeções são divididos em dois grupos: um grupo de treinamento e um grupo de teste. Cada método tenta fazer uma correlação entre as características de cada UC e o resultado das inspeções do grupo de treinamento, assim, "aprendendo"a identificar UC onde haja irregularidades ou a estimar uma probabilidade de se encontrar irregularidades em um grupo de UC analisadas. Uma vez terminada a etapa de treinamento, a capacidade de generalização da solução é testada aplicando o método de classificação no grupo de testes. Assim, as características desse grupo de teste são utilizadas como dados de entrada e a classificação obtida pelo método é comparada com os resultados obtidos em campo.

Em geral, o resultado obtido por um método de localização de PNT que realiza classificação de UC pode ser exibido a partir de uma matriz de confusão. Esta matriz compara os resultados da inspeção com os resultados de classificação de um método obtidos para um mesmo grupo de UC. Em geral, esse é o resultado da etapa de teste do método. Na posição "a"da matriz de confusão encontra-se a quantidade de UC corretamente classificadas como irregulares, também chamados de positivos verdadeiros. Na posição "b"encontra-se a quantidade de UC erradamente classificadas como irregulares, também conhecidos como falsos positivos. Na posição "c"é mostrada a quantidade de UC erradamente classificadas como regulares (falsos negativos) e na posição "d"tem-se a quantidade de UC corretamente classificadas como regulares, também conhecidos por negativos verdadeiros.

Tabela 3: Matriz de confusão.

		Rótulo correto	
		Irregular	Regular
Classificação	Irregular	a	b
	Regular	c	d

A partir dessa matriz é possível obter indicadores que avaliam a qualidade de cada método de classificação. Cada autor consultado utiliza uma nomenclatura diferente. Aqui essa nomenclatura será padronizada como mostrado a seguir:

- Valor preditivo positivo (VPP): Indica a fração dados corretamente classificados como irregulares pelo método de classificação com relação ao total de dados classificados como irregulares.

$$VPP = \frac{a}{a+b} \tag{8}$$

Também é conhecido como confiabilidade negativa (QUEIROGA, 2005). É um indicador importante, pois estima qual seria a proporção de UC com alguma irregularidade encontradas em inspeções realizadas a partir da classificação de clientes obtida por um determinado método de classificação. É o análogo à taxa de acerto de inspeções a clientes suspeitos.

- Total de classificação correta (TCC): Indica a quantidade de dados classificados corretamente.

$$TCC = \frac{a+d}{a+b+c+d} \tag{9}$$

Não é um indicador significativo para esta aplicação, pois o objetivo dos métodos de classificação é identificar UC com irregularidades para que as PNT sejam reduzidas de através de inspeções da forma mais eficiente possível, o que é obtido com altos índices VPP e $SENS$.

- Sensibilidade ($SENS$): É o número de UC irregulares classificadas corretamente sobre o número total de UC irregulares.

$$SENS = \frac{a}{a+c} \tag{10}$$

Indica a capacidade de incluir as UC irregularidades na categoria "irregular". Um método com índice $SENS$ alto é capaz de incluir na lista de UC a inspecionar a maior parte das UC irregulares, reduzindo, assim, a quantidade global de PNT.

- Precisão(Li): Enquanto a precisão, tal qual é apresentada em (11), é a medida de avaliação de desempenho de (PAPA et al., 2007) para o caso especial em que há apenas 2 classes.

$$Li = 1 - \frac{1}{2} \cdot \left(\frac{b}{c+d} + \frac{c}{a+b} \right) \tag{11}$$

Ela possui a vantagem de ponderar pelo tamanho de cada conjunto. Isso é desejável para aplicações como a de identificação de PC já que neste caso o número de elementos de uma das classes (irregular) é muito menor do que o da outra (normal) e o Li é uma pontuação que é ponderada pela taxa de acerto nas duas classes. No Brasil, por exemplo, as PC correspondem a aproximadamente 8% do consumo total. Em um caso assim, um TCC de 92% poderia parecer um resultado interessante, mas se os 8% de erro correspondessem aos elementos da classe "irregular", então, apesar do alto valor de TCC, o desempenho poderia ser considerado péssimo já que a classe de interesse foi aquela em que houve maior erro.

- F-score: O F-score é a média harmônica dos indicadores VPP e $SENS$. Dentre os indicadores listados é aquele que melhor se adapta à aplicação de classes desbalanceadas em que se tem o objetivo de obter o classificador que maximiza simultaneamente os indicadores $SENS$ e VPP.

$$F = \left(\frac{1}{VPP} + \frac{1}{SENS} \right)^{-1} \tag{12}$$

3.5 Conclusão

É grande o número de trabalhos realizados sobre o tema, inclusive alguns de companhias elétricas brasileiras. Dentre as metodologias para detecção e identificação de PC encontradas na literatura, destacam-se aquelas de localização de PC baseadas em IA. Estes métodos tem como objetivo a obtenção de uma lista de UC onde há grande probabilidade de se encontrar uma UC irregular. Dentre os quais três classificadores são utilizados em trabalhos cujos resultados são promissores: ANN, SVM e OPF. Logo, seria uma escolha sensata escolher uma dessas técnicas.

Dentre esses métodos, o que possui características mais interessantes é o OPF. É um método implicitamente multi-classes , o que significa que, apesar de ser aplicado a um problema de duas classes, ele é capaz de identificar diversos sub-padrões para PNT. Uma outra característica interessante é o tempo utilizado para treinamento, que, utilizando as implementações de (RAMOS et al., 2009) é muito superior às técnicas de RNA e SVM testadas. Além disso, é um método relativamente recente e que ainda precisa ser explorado.

Em geral, a maior parte dos métodos de classificação utilizam como dados de entrada perfis de carga diários ou histórico do consumo de UC. Outros indicadores, principalmente a localização geográfica, podem melhorar o desempenho dos classificadores. A investigação sobre os tipos de dados que afetam o desempenho de classificadores foi um dos dados ainda menos abordados no que diz respeito à localização de PNT e da investigação deste tema poderia surgir resultados interessantes. Os problemas relacionados aos dados cadastrais de UC nas concessionárias é que em geral ocorrem erros de cadastro e nem todas as distribuidoras possuem amplas bases de dados de seus clientes, portanto é importante buscar dados de entrada que sejam facilmente obtidos.

4 CLASSIFICADOR OPF

4.1 Introdução

O OPF é um *framework* para a o desenvolvimento de classificadores supervisionados e não supervisionados criado em 2005. A base destes métodos é a representação dos conjuntos de amostras como grafos. Os vetores de características, que são a representação matemática de cada amostra no espaço de amostras, são modelados como os nós do grafo enquanto as arestas representam as relações de adjacência entre cada nó a elas conectados. A aplicação inicial do OPF foi o reconhecimento de padrões em imagens, modelado como um problema de partição de um grafo. A estratégia empregada na resolução deste problema é fazer a segmentação do grafo em dois ou mais, de modo que as amostras de cada classe de dados fiquem em suas respectivas partições. Isso é realizado agrupando os nós do grafo em árvores, de acordo com regras diferentes para cada versão do OPF. Existem três versões: a versão com treinamento não supervisionado, a versão com treinamento supervisionado usando grafo k-NN e a versão com treinamento supervisionado usando grafo completo (figura 4 (a)). A última é a versão original do OPF e é considerada a que tem melhor desempenho, portanto é a que será abordada neste trabalho. Para tornar melhor a leitura, algumas definições sobre grafos não serão apresentadas ao longo do texto e estão contidas no Anexo A.

O classificador OPF na sua forma supervisionada é implementado em duas etapas fundamentais: a obtenção de protótipos e o treinamento do classificador por meio do algoritmo do OPF (PAPA et al., 2007). A primeira etapa tem como objetivo identificar quais amostras do grupo de treinamento (aqui chamado Z_1) são as mais relevantes, ou seja, aquelas que fornecem a melhor representação de cada classe de amostras. Estas amostras são chamadas de *protótipos*, que formam um conjunto representado pela letra S. A segunda etapa utiliza estas amostras para construir um grafo do tipo floresta de caminhos ótimos (figura 4 (b)), cujas raízes são os protótipos encontrados anteriormente, agrupando as amostras próximas às raízes em subgrafos do tipo árvore. Além dessas etapas fundamentais, podem ser implementados algoritmos para melhoria de desempenho do classificador. Alguns desses algoritmos necessitam de um segundo conjunto para o projeto do classificador, chamado de conjunto de aprendizado ou de validação cruzada (Z_2), dependendo do algoritmo em questão.

O teste do classificador projetado é realizado sobre um conjunto de dados rotulados, chamado Z_3. Cada amostra é inserida no grafo floresta de caminhos ótimos obtido no treinamento, conectando-se por arestas a todos os elementos do conjunto de treinamento (figura 4 (c)). A árvore que oferece o menor custo de caminho à amostra classificada à conquista para sua classe (figura 4 (d)).

Na fase de testes ou na fase de classificação, uma amostra por vez é inserida no grafo

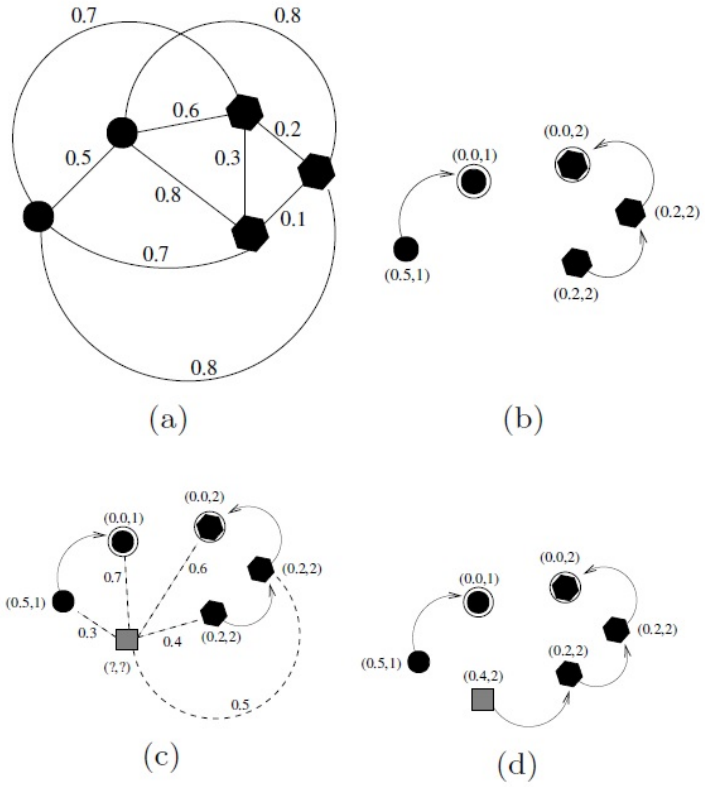

Figura 4: Etapas de treinamento e classificação de um classificador OPF. a) grafo obtido com o conjunto de teste. b) geração de floresta de caminhos ótimos. c) inserção de uma amostra no grafo depois do treinamento. d) classificação da nova amostra na classe "hexágono". Adaptado de (PAPA et al., 2007).

junto com a floresta e é verificado então qual das árvores a conquistou, classificando a amostra (PAPA; FALCÃO; SUZUKI, 2009). A amostra então é retirada do grafo e a classificação da amostra seguinte pode ser realizada.

As amostras são representadas em um espaço de características por um vetor de características v, de n dimensões. Este vetor de características representa a posição de cada nó do grafo em um espaço n-dimensional. No caso do problema de identificação de perdas comerciais através da análise de curvas de carga, estas características podem ser valores numéricos como, por exemplo, consumo instantâneo, consumo médio diário, consumo mensal, ou desvio padrão do consumo diário ou ainda outros tipos de dados de UCs como localização da UC (rua, bairro, etc), transformador a que está conectado, classe de consumo, atividade econômica, entre outros. A métrica para definir a distância entre duas amostras é a distância euclidiana d entre os vetores de características de cada amostra. Pode-se adotar outras métricas de distância entre amostras diferentes da euclidiana. Essa distância é denotada $d(s,t)$ para a distância entre uma amostra s e uma amostra t.

O problema consiste em utilizar S, v, d, Z_1 e Z_2 para projetar um classificador ótimo que seja capaz de prever o rótulo correto de qualquer amostra pertencente ao conjunto de teste Z_3. O classificador cria uma partição ótima no espaço de características, que é uma floresta de caminhos ótimos, calculada em $\Re^n$ pela algoritmo de Transformada Imagem Floresta (PAPA; FALCÃO; SUZUKI, 2009).

4.2 Obtenção de Protótipos

Os protótipos são aquelas amostras que fornecem a melhor representação para cada padrão. Essas amostras funcionam de maneira análoga aos *support vectors* do classificador SVM. Os protótipos ótimos de um grafo são obtidos por meio da relação teórica entre os grafos de árvore geradora mínima e árvore de caminhos mínimos (PAPA et al., 2007). Essa relação garante que os protótipos ótimos são os vértices conectados a vértices de outras classes pelas arestas da árvore geradora mínima. Logo, para obter os protótipos ótimos do grupo de treinamento, deve-se computar a sua árvore geradora mínima e analisar a conexão entre seus nós de diferentes classes. A árvore geradora mínima é um subgrafo que conecta todos os nós do grafo original e que possui $n - 1$ arcos, onde n é o número de nós. Os vértices da árvore geradora mínima que são conectados a elementos de classes diferentes formam o conjunto dos protótipos, representado por S. É necessário que haja pelo menos um protótipo por classe.

Para obtenção da árvore geradora mínima podem ser utilizados vários algoritmos, como o algoritmo guloso de Kruskal (algoritmo 5, página 88) e o algoritmo de Prim (algoritmo 6, página 89), por exemplo (KREYSZIG, 1997). Esse algoritmo garante que a árvore geradora encontrada é mínima. A escolha do algoritmo a ser utilizado depende do problema ao qual eles são aplicados. Em geral, o algoritmo de Kruskal é mais adaptado para grafos esparsos, que são aqueles que possuem poucas arestas, já o algoritmo de Prim é mais adequado para o caso oposto, quando há grafos densos. Uma vantagem do algoritmo de Kruskal é a sua maior simplicidade de implementação.

4.3 Treinamento

Depois de obter o conjunto de protótipos S pode-se realizar o treinamento do classificador por meio do algoritmo de Floresta de Caminhos Ótimos, que chamaremos de OPF. O objetivo deste algoritmo é obter, a partir de um grafo completo, uma floresta de caminhos de custo mínimo que, por sua vez, são enraizadas nos nós protótipos e possuem apenas elementos de uma classe de Z_1. Alternativamente, pode-se pensar que os protótipos "aglutinam"ou "conquistam"os nós mais próximos a eles, formando conglomerados, que são utilizados na etapa de classificação para conquistar as amostras não-rotuladas (de classe desconhecida) mais próximas a eles, de acordo com algum critério. A conquista de amostras é o ato de projetar a classe de um protótipo em uma amostra ainda não classificada. A métrica para decisão dessas conquistas é uma função de custo de caminho suave que tem como entradas os pesos das arestas que ligam os protótipos a cada nó do grafo (PAPA; FALCÃO; SUZUKI, 2009). relacionada com os pesos dados pela distância de cada aresta, que pode ser uma função de distância euclidiana, por exemplo. Após todos os nós serem conquistados, cada protótipo dá origem a uma árvore de caminho ótimo. A presença de várias árvores, por conta da existência de vários protótipos, dá ideia de uma floresta, daí o nome do método. O critério utilizado na versão supervisionada do OPF para a conquista de amostras é a menor função de custo de caminho, chamada f_{max}. Ela é definida por (13).

$$\begin{aligned} f_{max}(\langle s \rangle) &= \begin{cases} 0 & se\ \mathrm{s} \in\ S \\ +\infty & caso\ contrário \end{cases} \\ f_{max}(\pi \cdot \langle s,t \rangle) &= max\{f_{max}(\pi), d(s,t)\} \end{aligned} \tag{13}$$

onde π é um caminho, s é uma amostra e $\pi \cdot \langle s \rangle$ é a concatenação do caminho π, que termina na amostra t e uma amostra s. Percebe-se que f_{max} possui características dife-

rentes quando avalia caminhos triviais (grafos do tipo caminho que contém apenas um nó) e quando avalia caminhos contendo mais de um elemento. Um caminho é dito *ótimo* quando sua função custo é mínima quando comparada a qualquer outro caminho que termina na mesma amostra.

O algoritmo OPF parte de um grafo completo do qual se conhecem os protótipos (figura 5 (a))e atribui um caminho ótimo $P^*(s)$ de S para cada amostra s pertencente a Z_1, formando uma floresta de caminhos ótimos P. Cada nó do grafo herda o rótulo do nó raiz do caminho ao qual pertencem. $P(s)$ contém o predecessor de cada nó no caminho, e recebe um marcador nulo quando o nó é uma raiz. $C(s)$ contém o custo de cada amostra, conforme calculado pelo algoritmo pela função f_{max}. A partir do conjunto de treinamento Z_1 e seus respectivos rótulos dados pela função $\lambda(s)$, do conjunto de protótipos S, das distâncias entre as amostas e os vetores de características de todos os elementos de Z_1, o algoritmo OPF obtém uma floresta de caminhos ótimos. Os resultados deste algoritmo são mapa de custos $C(s)$, que associa cada amostra a um custo de caminho dado pela função f_{max} e utilizado no cálculo de cada caminho ótimo, o mapa de rótulos $L(s)$ que associa cada elemento de Z_1 ao seu respectivo rótulo de tal maneira que $L(s) = \lambda(s)$ e a floresta de caminhos ótimos (figura 5 (b)) representada através da lista de precedência $P(s)$ para todas as amostras do conjunto de treinamento. A lista de precedência $P(s)$ indica qual vértice é o precedente ao vértice s no caminho que liga o vértice s à raiz da sua árvore. Como o nó-raiz por definição não possui vértice precedente, será utilizado o marcador nil neste caso. São usados também a fila de prioridades Q, da qual saem primeiro as amostras com menor custo, e a variável cst é utilizada para se atualizar o mapa de custos das amostras, caso seja encontrado para a amostra em questão um caminho cujo custo é menor que aquele em que a amostra se encontra.

Nas linhas 1 a 3 do algoritmo ocorre a inicialização do mapa de custos, do mapa de rótulos e a floresta de caminhos ótimos. Para as amostras de Z_1 que não pertencem a S e, portanto ainda não foram rotuladas, o custo é $+\infty$, conforme é definido na linha 1. Para os protótipos, o custo $C(s)$ tal que $s \in S$ é nulo, o que é definido na linha 3. Na mesma linha são inicializados o mapa de rótulos e os predecessores dos protótipos. Além disso, todos os protótipos são inseridos na fila Q. O laço executado nas linhas 4 a 10 só termina quando se esvazia a fila e o seu objetivo é formar a floresta de caminhos ótimos alocando cada amostra no seu caminho ótimo, que é aquele de menor custo possível. Na linha 5, retira-se da fila a amostra de menor custo, chamada s. Se existe em Z_1 alguma outra amostra t cujo custo é maior que o custo de s, é calculado o valor cst, que é o maior valor entre o custo de s e a distância de s a t, na linha 7 e se este custo for menor do que o custo $C(t)$, a amostra t é inserida em outro caminho. Caso a amostra já tenha sido rotulada antes (linha 9), ela precisa ser removida da lista de prioridades para ser novamente incluída na linha seguinte. Caso contrário, trata-se de uma amostra que não está na lista de prioridades Q e é então ali inserida. Finalmente, atualizam-se o rótulo, o predecessor e o custo de t de acordo com a amostra s que a "conquistou", na linha 10. O algoritmo para quando a última amostra de Q é retirada e não há mais amostras em Z_1 cujo custo é superior ao desta.

Algoritmo 1: Floresta de Caminhos Ótimos

Entrada: Conjunto de treinamento Z_1, cujos rótulos das amostras são λ; conjunto de protótipos $S \subset Z_1$; vetor de características (descrição dos nós do grafo) e distâncias entre amostras.

Saída: Floresta de Caminhos Ótimos P, mapa de Custo C e mapa de rótulos L.

Auxiliar: Fila de prioridades Q e variável de custo cst.

1. **para cada** $s \in Z_1 \backslash S$ **faça** $C(s) \leftarrow +\infty$
2. **para cada** $s \in S$ **faça**
3. $\quad C(s) \leftarrow 0$, $P(s) \leftarrow nil$, $L(s) \leftarrow \lambda(s)$ e inserir s em Q.
4. **enquanto** Q *não está vazia* **faça**
5. $\quad$ Retirar de Q uma amostra s tal que seu $C(s)$ é mínimo.
6. $\quad$ **para cada** $t \in Z_1$ *tal que* $t \neq s$ *e* $C(t) > C(s)$ **faça**
7. $\quad\quad$ Calcule $cst \leftarrow max\{C(s), d(s,t)\}$
8. $\quad\quad$ **se** $cst < C(t)$ **então**
9. $\quad\quad\quad$ **se** $C(t) \neq +\infty$ **então** remova t de Q.
10. $\quad\quad\quad$ $P(t) \leftarrow s$, $L(t) \leftarrow L(s)$, $C(t) \leftarrow cst$ e insira t em Q.

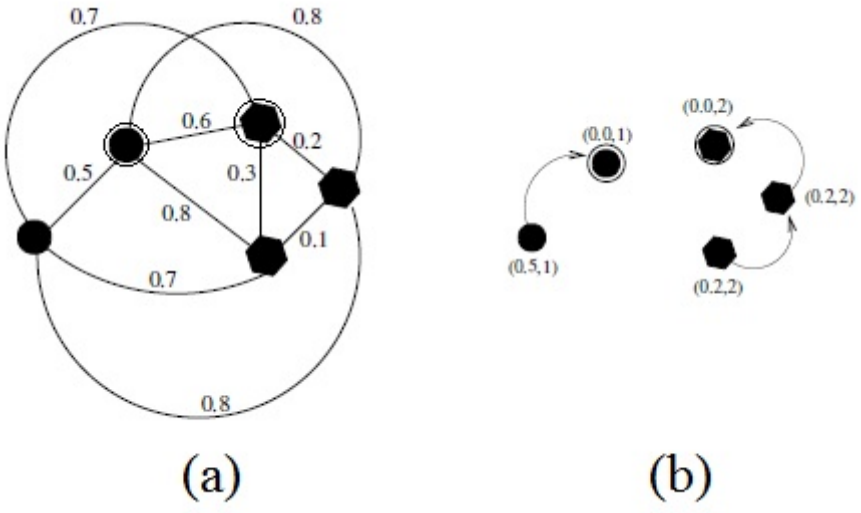

Figura 5: Em (a) tem-se o grafo completo contendo os protótipos (nós circulados), o ponto de partida do algoritmo OPF. Em (b) temos a floresta de caminhos ótimos, mostrando o classificador treinado. Entre parênteses tem-se o par $(C(s), L(s))$ para cada nó. Adaptado de (PAPA et al., 2007).

4.4 Classificação

A classificação dos elementos do conjunto de teste é feita elemento a elemento, inserindo uma amostra no grafo floresta de caminhos ótimos. Esta amostra é conectada por uma aresta a cada elemento da floresta de caminhos ótimos, portanto, a cada elemento do conjunto Z_1 (figura 6 (a)). Deseja-se então descobrir a qual protótipo esta amostra é mais fortemente conectada. Para isso recorremos outra vez à função f_{max}, calculando o seu valor para cada elemento da floresta de caminhos ótimos obtida na etapa de treinamento. Escolhe-se então aquela que possui menor valor de f_{max}, que conquistará a amostra classificada, passando a esta seu rótulo. O valor de f_{max} em questão será o valor da função custo da nova amostra. De maneira mais formal, deve-se utilizar (14) para obter o custo da amostra t a classificar, isto é, dar a esta o rótulo da amostra s cujo f_{max} é mínimo (figura 6 (b)). Depois que uma amostra é classificada, ela é retirada do grafo e o processo se repete para a amostra seguinte.

$$C(t) = min\{max\{C(s), d(s,t)\}\}, \forall s \in Z_1 \tag{14}$$

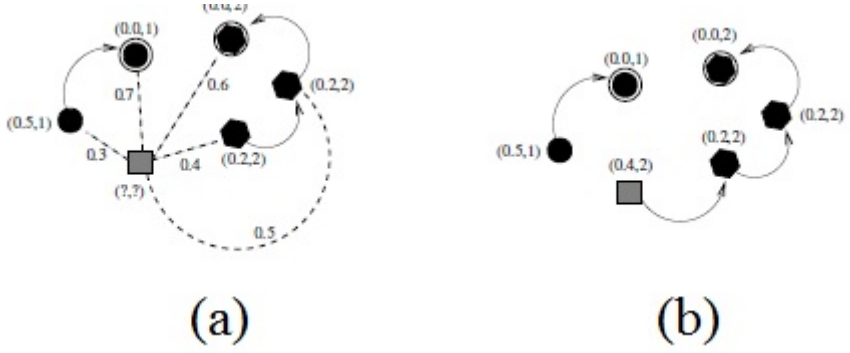

Figura 6: Exemplo de classificação de uma amostra pelo OPF. Em (a) tem-se o grafo formado pela inserção da nova amostra, ainda não rotulada, e todos os nós da floresta de caminhos ótimos obtida na fase de treinamento. Em (b) é mostrado que a amostra foi classificada na classe dos nós hexagonais ("2"). Adaptado de (PAPA et al., 2007).

A etapa de classificação do OPF é um dos seus pontos fracos no que diz respeito ao tempo de execução do algoritmo. Como para cada amostra do conjunto de treinamento é necessário calcular o peso de tantas arestas quantos são os elementos de Z_1 e é necessário avaliar o caminho de custo mínimo para cada amostra, a operação é relativamente lenta quando a comparamos a outros algoritmos e dependente do número de amostras do conjunto de treinamento além do número de elementos do conjunto de teste.

4.5 Melhoria do Desempenho

4.5.1 Aprendizado

Quando trabalha-se com grandes quantidades de dados, o tempo de execução dos algoritmos pode tornar-se um problema. Bases de dados reais de distribuidoras de energia elétrica possuem de milhares a milhões de clientes cadastrados. Sabe-se que, por exemplo, o algoritmo de Kruskal para se encontrar árvores geradoras mínimas, utilizado para encontrar os protótipos do OPF, possui complexidade $O\left(m^2\right)$, onde m é o número de arestas do grafo. Então, o número de operações executadas é proporcional ao quadrado

do número de elementos. Durante a classificação, há um problema parecido, já que há uma etapa de ordenamento, cuja complexidade é $O\left(m \cdot log\left(m\right)\right)$. Logo, para que a execução do classificador não seja demasiado lenta, é interessante estudar maneiras de se diminuir o número de amostras necessárias para se projetar um classificador com desempenho desejado. Uma maneira de abordar esse problema é dividir o grupo de treinamento em dois: um grupo de treinamento propriamente dito Z_1 e um grupo de avaliação Z_2, cuja função é enviar ao grupo Z_1 as amostras mais representativas. Então, o algoritmo de aprendizado, a partir desses dois conjuntos, define um novo conjunto com o mesmo número de amostras de Z_1 cujo classificador é aquele com maior precisão possível para aqueles dois conjuntos. O algoritmo 2 é encontrado na referência (PAPA et al., 2007). O algoritmo utilizado para o projeto do classificador do OPF é o mesmo apresentado na seção 4.3. O critério de classificação também é o mesmo, dado por (14).

Define-se então s^* a amostra de Z_1 que conquistou a amostra a classificar t, satisfazendo a equação (14). Projeta-se um classificador utilizando Z_1 e as amostras de Z_2 são classificadas pelo mesmo. O princípio deste algoritmo é fazer a substituição de amostras irrelevantes de Z_1 por aquelas consideradas mais informativas em Z_2. As amostras de Z_1 consideradas irrelevantes são aquelas que fazem parte de um caminho que erra mais do que acerta ao classificar amostras de Z_2. Estes nós são colocados na lista de nós irrelevantes (LI). Além disso, considera-se que as amostras mais informativas de Z_2 são aquelas que foram classificadas erradamente. São contados, a cada iteração, o número de amostras de Z_2 classificadas corretamente e incorretamente por cada elemento de Z_1 e todos os nós cujo caminho termina neste nó (da raiz até o nó que conquistou a amostra que está sendo classificada). Os nós de Z_2 classificados incorretamente são colocados em uma lista de nós classificados incorretamente (LE). Em um primeiro momento, são substituídas as amostras de Z_2 classificadas erradamente pelas amostras de Z_1 consideradas irrelevantes. Os demais nós irrelevantes de Z_1 são substituídos por nós de Z_2, desde que estes não tenham pertencido a Z_1 anteriormente.

As operações de classificação e substituição se repetem por um número arbitrário de iterações. Para evitar que as amostras troquem de grupo a cada iteração, um vetor auxiliar chamado TR lista as amostras que foram de Z_2 para Z_1 e essas amostras listadas são impedidas de fazerem o caminho de volta. Os números de falsos positivos e falsos negativos são contabilizados para facilitar a avaliação da precisão do classificador. Os vetores $NR(s)$ e $NW(s)$ contabilizam o número de vezes que uma determinada amostra s pertencente a Z_1 fez parte do caminho ótimo que classificou uma amostra de Z_2 corretamente ou incorretamente, respectivamente. A função $\lambda(s)$ fornece o rótulo correto da amostra s.

As amostras de Z_1 consideradas irrelevantes são armazenadas na lista LI enquanto as amostras de Z_2 classificadas incorretamente são armazenadas em LE. Nas linhas 1 a 7 ocorre a inicialização dos vetores auxiliares e o treinamento do OPF. O único vetor auxiliar que não é atualizado a cada iteração é TR. Nas linhas 9 a 18 os elementos de Z_2 classificados corretamente ou incorretamente em são contabilizados, bem como os acertos e erros dos elementos do classificador. Nas linhas 20 a 22 são contabilizadas as amostras irrelevantes. Nas linhas 24 a 26 ocorre a substituição dos elementos irrelevantes de Z_1 pelos elementos classificados incorretamente de Z_2. Nas linhas 27 a 29 elementos considerados irrelevantes de Z_1 são substituídos por elementos de Z_2 escolhidos ao acaso mas que não pertençam à lista TR. As trocas de amostras ocorreram apenas entre elementos pertencentes às mesmas classes.

Uma versão modificada do algoritmo de aprendizado (algoritmo 2) é apresentada apre-

sentada em (RAMOS et al., 2011), o algoritmo 3, uma versão aplicada a PNT. Nesta versão, a cada iteração, amostras de Z_2 classificadas incorretamente são incluídas em uma lista, chamada LM (*List of Misclassified samples*). Elementos dessa lista são trocados por amostras de Z_1 escolhidas aleatoriamenteque não sejam protótipos. O melhor classificador, aquele com maior precisão, é salvo.

4.5.2 Poda

O algoritmo de poda, como é mostrado em (RAMOS et al., 2011), tem como objetivo reduzir o tamanho do conjunto de treinamento, de maneira a reduzir o tempo necessário para realizar as etapas de treinamento e classificação. Isso é realizado removendo amostras irrelevantes do projeto do classificador. Amostras ditas irrelevantes, neste caso, são aquelas que não são utilizadas para classificar nenhuma amostra do conjunto de avaliação Z_2. Este algoritmo tem como ponto de partida o classificador obtido com o algoritmo de aprendizado do OPF. Aplicando esse classificador às amostras de Z_2, pode-se determinar quais amostras do conjunto de avaliação foram conquistadas por quais nós. Se um nó de uma árvore conquista uma ou mais amostras, todos os elementos do caminho ótimo existente entre esse nó e a raiz, incluindo estes dois, são considerados relevantes (Linhas 6 a 8). As amostras irrelevantes de Z_1 são passadas para o conjunto Z_2 (linha 10). Mesmo as amostras que classificam erradamente a base de dados são classificadas como relevantes.

Uma melhoria possível nestes algoritmos é a substituição do indicador Li por algum outro mais pertinente para a aplicação desejada, como o $F - score$ ou VPP.

Algoritmo 2: Aprendizado do OPF.

Entrada: Conjuntos de treinamento Z_1 e avaliação Z_2, cujos rótulos das amostras são λ; conjunto de protótipos $S \subset Z_1$; vetor de características, distâncias entre amostras e número de iterações T.

Saída: Curva de aprendizado L e o último classificador OPF obtido.

Auxiliar: Vetores FP e FN de tamanhos c para falsos positivos e falsos negativos, listas LI e LE de amostras irrelevantes e amostras classificadas erradamente para cada classe, vetores NR e NW de tamanho $|Z_1|$ contabilizando classificações corretas e incorretas, variáveis s para amostras de Z_1, t para amostras de Z_2, e r para antecessor na árvore e conjunto TR para evitar que amostras de Z_2 retornem para Z_1.

1. $TR \leftarrow \emptyset$
2. **para** *cada iteração* $I = 1, 2, 3, ..., T$ **faça**
3. $\quad TR \leftarrow TR \cup Z_1$
4. $\quad$ Treinar OPF com Z_1
5. $\quad$ **para cada** *amostra* $t \in Z_2$ **faça** $NR(2) \leftarrow 0$ e $NW(s) \leftarrow 0$
6. $\quad$ **para cada** *classe* $i = 1, 2, 3, ..., c$ **faça**
7. $\quad\quad FP(i) \leftarrow 0, FN(i) \leftarrow 0, LI \leftarrow \emptyset$ e $LE \leftarrow \emptyset$
8. $\quad$ **para cada** *amostra* $t \in Z_2$ **faça**
9. $\quad\quad$ Encontrar $s^* \in Z_1$ que satisfaz a equação (14) e fazer $r \leftarrow s^*$
10. $\quad\quad$ **se** $L(t) \neq \lambda(t)$ **então**
11. $\quad\quad\quad FP(L(s^*)) \leftarrow FP(L(s^*)) + 1$
12. $\quad\quad\quad FN(\lambda(s^*)) \leftarrow FN(\lambda(s^*)) + 1$
13. $\quad\quad\quad$ **se** $t \notin TR$ **então** $LE(\lambda(t)) \leftarrow LE(\lambda(t)) \cup \{t\}$
14. $\quad\quad\quad$ **enquanto** $r \neq nil$ **faça**
15. $\quad\quad\quad\quad NW(r) \leftarrow NW(r) + 1$ *e* $r \leftarrow P(r)$
16. $\quad\quad$ **senão**
17. $\quad\quad\quad$ **enquanto** $r \neq nil$ **faça**
18. $\quad\quad\quad\quad NR(r) \leftarrow NR(r) + 1$ *e* $r \leftarrow P(r)$
19. $\quad$ Calcular precisão $Li(I)$
20. $\quad$ **para cada** *amostra* $s \in Z_1$ **faça**
21. $\quad\quad$ **se** $NW(s) > NR(s)$ **então**
22. $\quad\quad\quad LI(\lambda(s)) \leftarrow LI(\lambda(s)) \cup \{s\}$
23. $\quad$ **para cada** *classe* $i = 1, 2, 3, ..., c$ **faça**
24. $\quad\quad$ **enquanto** $|LI(i)| > 0$ *e* $|LE(i)| > 0$ **faça**
25. $\quad\quad\quad LI(i) \leftarrow LI(i) \backslash \{s\}$ e $LE(i) \leftarrow LE(i) \backslash \{t\}$
26. $\quad\quad\quad$ Substituir $s \in Z_1$ por $t \in Z_2$
27. $\quad\quad$ **enquanto** $|LI(i)| > 0$ **faça**
28. $\quad\quad\quad LI(i) \leftarrow LI(i) \backslash \{s\}$
29. $\quad\quad\quad$ Encontrar $t \in Z_2 \backslash TR$, com $\lambda(i) = i$ e substitui-lo por $s \in Z_2$
30. Treinar OPF com Z_1

Algoritmo 3: Aprendizado do OPF para perdas não-técnicas.

Entrada: Conjuntos de treinamento Z_1 e avaliação Z_2, cujos rótulos das amostras são λ; conjunto de protótipos $S \subset Z_1$; vetor de características, distâncias entre amostras e número de iterações T.

Saída: Curva de aprendizado L e o último classificador OPF obtido.

Auxiliar: Vetores FP e FN de tamanhos c (nº de classes) para falsos positivos e falsos negativos e lista LM para amostras classificadas incorretamente.

1. Definir $MaxAcc = -1$
2. **para** *cada iteração* $I = 1, 2, 3, ..., T$ **faça**
3. $\quad LM = 0$
4. $\quad$ Treinar OPF com Z_1
5. $\quad$ **para cada** *classe* $i = 1, 2, 3, ..., c$ **faça**
6. $\quad\quad FP(i) \leftarrow 0, FN(i) \leftarrow 0$
7. $\quad$ **para cada** *amostra* $t \in Z_2$ **faça**
8. $\quad\quad$ Utilizar classificador obtido na linha 4 para classificar com rótulo $L(t)$
9. $\quad\quad$ **se** $L(t) \neq \lambda(t)$ **então**
10. $\quad\quad\quad FP(L(t)) \leftarrow FP(L(t)) + 1$
11. $\quad\quad\quad FN(\lambda(t)) \leftarrow FN(\lambda(t)) + 1$
12. $\quad\quad\quad LM \leftarrow LM \cup \{t\}$
13. $\quad$ Calcular precisão $Li(I)$
14. $\quad$ **se** $Li(I) > MaxAcc$ **então** salvar classificador atual
15. $\quad MaxAcc \leftarrow Li(I)$
16. $\quad$ **enquanto** $LM \neq \emptyset$ **faça**
17. $\quad\quad LM \leftarrow LM \backslash t$
18. $\quad\quad$ Substituir t por $s \in Z_1$ aleatório tal que $\lambda(s) = \lambda(t)$ e $s \notin S$

Algoritmo 4: Aprendizado com poda para perdas não-técnicas

Entrada: Conjuntos de treinamento Z_1 e avaliação Z_2, cujos rótulos das amostras são λ; conjunto de protótipos $S \subset Z_1$; vetor de características, distâncias entre amostras e número de iterações T.

Saída: Classificador OPF com conjunto de treinamento Z_1 reduzido.

Auxiliar: Conjuntos R e I de amostras relevantes e irrelevantes, respectivamente.

1. **enquanto** $I \neq \emptyset$ **faça**
2. $\quad$ Executar algoritmo de aprendizado do OPF
3. $\quad R \leftarrow \emptyset, I \leftarrow \emptyset$
4. $\quad$ **para cada** *amostra* $t \in Z_2$ **faça**
5. $\quad\quad$ Classificar t
6. $\quad\quad$ **enquanto** $P(t) \neq nil$ **faça**
7. $\quad\quad\quad$ Inserir $P(t)$ em R
8. $\quad\quad\quad t \leftarrow P(t)$
9. $\quad I \leftarrow Z_1 \backslash R$
10. $\quad$ Mover amostras de $Z_1 \backslash R$ para Z_2

4.6 Conclusão

Neste capítulo foram apresentados os aspectos fundamentais do classificador OPF em sua versão supervisionada, além de alguns algoritmos para melhoria de desempenho. Nesta versão, são utilizados três conjuntos de dados rotulados (ou seja, dados cuja classe é conhecida de antemão) para o projeto e a avaliação do classificador. O primeiro deles, chamado de conjunto de treinamento (Z_1), é utilizado para o projeto do classificador. O segundo é dito de aprendizado (Z_2) é utilizado para o projeto somente quando se deseja realizar etapas de melhoria de desempenho como aprendizado e poda (algoritmos 2, 3 e 4). O terceiro deles, Z_3 é o conjunto de teste, utilizado para testar o desempenho do classificador obtido a partir de Z_1 e Z_2. A primeira etapa do projeto do classificador é a identificação das amostras mais significativas de cada classe, chamadas de protótipos. Eles são os elementos de Z_1 conectados a elementos de classes diferentes da sua pela árvore geradora mínima obtida a partir do conjunto de treinamento. A partir dos protótipos obtidos realiza-se segunda etapa do treinamento que é a obtenção da floresta de caminhos ótimos pelos algoritmo do OPF (algoritmo 1).

Uma vez obtido o classificador, pode-se utilizar algoritmos para melhora do desempenho como os de aprendizado, que melhora o desempenho do classificador ao substituir amostras irrelevantes de Z_1 por amostras relevantes de Z_2 ou associar o aprendizado à poda, de maneira a reduzir o tempo de treinamento e classificação por meio da diminuição do tamanho de Z_1, sem que haja comprometimento do desempenho do classificador. Após obtido o classificador OPF, pode-se então avaliar seu desempenho na etapa de classificação do conjunto Z_3. A classificação do OPF é realizada inserindo-se uma amostra do conjunto de classificação no grafo OPF, conectando essa amostra a todos os elementos de Z_1. A classificação é dada calculando a menor função de custo de caminho mínimo (equação (14)). O caminho existente na floresta de caminhos ótimos que fornece o menor custo à amostra sendo classificada "conquista"esta amostra e fornece a ela o seu rótulo, classificando a mesma.

O objetivo geral do OPF é realizar a partição de zonas do espaço de características para cada classe de elementos disponível. Essas partições são delimitadas pela distância entre as árvores geradas pelo algoritmo OPF. Uma das vantagens do OPF é que essas regiões são definidas implicitamente e podem facilmente assumir formatos complexos, diferentemente de outros classificadores como SVM que necessitam de *kernels* e consequente aumento do número de dimensões (como o SVM-RBF) para que as regiões assumam formatos complexos. Outra vantagem é que, como o número de árvores na floresta de caminhos ótimos é tão grande quanto o número de protótipos, o classificador é implicitamente multi-classes. Em mais uma comparação com o SVM, este possui em sua versão linear, por exemplo, apenas duas classes já que o SVM linear realiza a divisão do espaço de características em duas, por meio de um hiper-plano. Durante o treinamento de classificadores como o SVM ou RNA é necessário utilizar um método de otimização para ajuste de coeficientes. No caso do OPF, isso não é necessário, uma vez que não há parâmetros para ajustar. Esse é um fator que contribui para que a etapa de treinamento do OPF seja mais rápida do que a dos dois classificadores citados. Já na etapa de classificação encontramos o ponto fraco do OPF. Como a etapa é realizada amostra por amostra, ela é relativamente lenta. Em métodos como SVM linear e RNA, a classificação é feita por meio de uma ou mais multiplicações matriciais, que podem ser indexadas, resultando em um uma etapa de classificação que em geral é mais rápida do que aquela realizada pelo OPF. Tendo em vista os resultados superiores do OPF com relação a outros classificadores bastante reconhecidos relatados na literatura e as vantagens do OPF listadas

acima, tomou-se a decisão de empregar este classificador como método de escolha para a detecção e identificação de PNT.

5 CLASSIFICADOR OPF APLICADO A PERDAS COMERCIAIS

5.1 Introdução

Neste trabalho o problema de identificação de PC é tratado como um problema de reconhecimento de padrões. Logo, parte-se do pressuposto que a partir da análise dos dados disponíveis é possível ensinar um classificador a encontrar padrões de PC de maneira eficiente. Além da necessidade dessa premissa ser verdadeira, pelo menos em certo grau, e de um bom classificador para encontrar corretamente os padrões desejados, o sucesso desse tipo de sistema é altamente dependente da informação que ele recebe e como ela é pré-processada. As variáveis de entrada escolhidas e o volume de dados disponíveis são elementos críticos. Idealmente, o conjunto de dados a disposição para projeto do sistema de reconhecimento de padrões deve conter uma amostra representativa do espaço de características. Em suma, pode-se dizer de maneira mais formal que é necessário que os dados existam em quantidade suficiente e que as características utilizadas sejam processadas de maneira que as classes desejadas ocupem regiões diferentes do espaço de características e que o classificador seja capaz de fazer a separação correta, ou ao menos eficiente, destes espaços.

Neste capítulo, as particularidades do sistema de reconhecimento de padrões desenvolvido para a detecção e identificação de PC são apresentadas em detalhe. Uma visão geral do sistema proposto é mostrado na seção 5.2. Na seção 5.3 são apresentados os dados de entrada considerados, bem como a maneira que eles são pré-processados. O classificador utilizado é apresentado em 5.4. Na seção 5.5 são apresentados dados relativos ao pós-processamento dos dados produzidos pelo classificador OPF. Ao final do capítulo, uma conclusão é apresentada em 5.6.

5.2 Fluxograma

O sistema de detecção e identificação de PC proposto tem como objetivo reduzir a quantidade de PC por meio da identificação e sinalização dessas perdas para que, posteriormente, possam ser realizadas ações para que elas sejam eliminadas. As inspeções realizadas nas UC para verificação de irregularidades são as ações mais importantes neste sentido. Uma vez constatada a irregularidade, o prejuízo da concessionária de energia elétrica pode ser revertido e a UC pode ser regularizada, eliminando a PC naquele local. Como esse tipo de ação gera custos, além de poder gerar constrangimento aos clientes inspecionados, é muito importante que a taxa de acerto do sistema seja a mais alta possível.

O processo de identificação de PC realizado pelo sistema proposto é resumido no fluxograma mostrado na figura 7. Com base nos dados de cadastro dos clientes e do seu histórico mensal de consumo, um algoritmo de extração de características acessa essa base de dados e transforma as informações da base de dados em vetores de características. Uma filtragem desses dados é realizada para eliminação de dados incoerentes. O classificador OPF é então treinado utilizando os dados das amostras rotuladas presentes na base de dados. As amostras rotuladas são as informações relativas às UC já inspecionadas anteriormente, cuja classe já é conhecida a princípio. Uma vez treinado, o classificador OPF realiza a classificação daquelas amostras não rotuladas e gera uma lista de UC candidatas à inspeção. Uma etapa de pós-processamento é realizada para o tratamento dos resultados do classificador. Uma vez que a lista resultante do pós-processamento é gerada, as equipes de inspeção são enviadas aos locais apontados. Depois de realizadas as inspeções, os seus resultados contendo a confirmação ou não da suspeita são armazenados na base de dados da empresa.

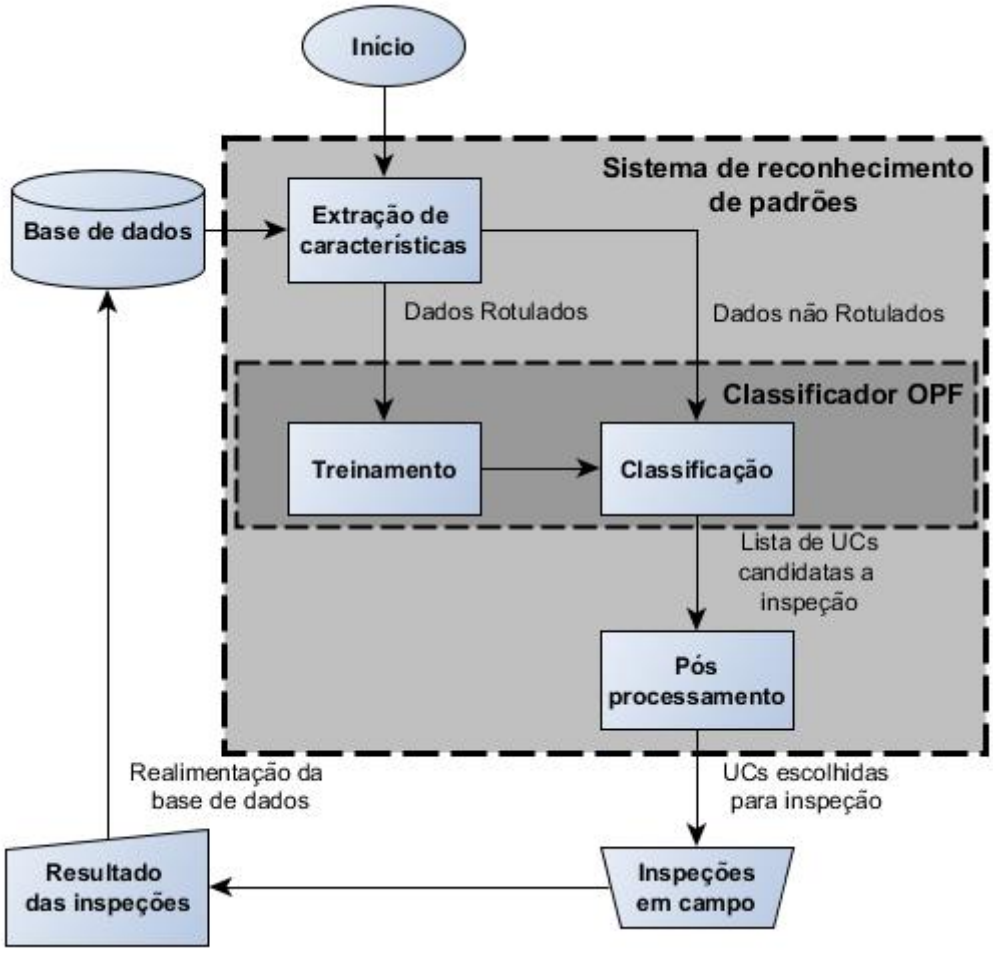

Figura 7: Fluxograma do sistema de reconhecimento de padrões proposto. Fonte: Próprio autor.

5.3 Dados de Entrada

Conforme foi citado em detalhe na seção 3.3 empresas distribuidoras de energia elétrica mantêm bases de dados contendo grandes quantidades de dados de clientes. No entanto, entrando em contato com algumas empresas distribuidoras, ficou evidente que

o conteúdo dessas bases não é padronizado e as informações contidas nelas podem ser bastante heterogêneas. Portanto, tomou-se a decisão de trabalhar com informações de cadastro essenciais, tais como:

- históricos mensais de consumo;
- endereço;
- transformador de conexão;
- classe de consumo;

Pode-se considerar que, à exceção dos históricos de consumo, os demais dados são categóricos, ou seja, não possuem valor numérico. Logo, para que estes possam ser utilizados juntamente com o OPF, é necessário uma etapa de tratamento desses dados para que eles assumam valores numéricos adequados à aplicação.

5.3.1 Histórico Mensal de Consumo

De acordo com o que é citado na seção 3.3 as distribuidoras são obrigadas por normativa da ANEEL a armazenar dados de histórico mensal de consumo. Considerando que as PC caracterizam-se como energia não faturada, que tem como consequência perdas financeiras para as concessionárias de energia elétrica, uma variável muito importante para o estudo é o histórico mensal de consumo. Para o projeto dos classificadores e como filosofia geral do sistema de reconhecimento de padrões proposto, foi postulado que quando surge uma PC em uma UC, como o aparecimento do defeito, alteração ou fraude de medidor e até derivações na instalação elétrica, o consumo mensal da UC tende a diminuir. Para destacar essa mudança repentina de consumo e poder identificar esta característica em históricos com diferentes magnitudes de consumo mensal, utilizou-se da normalização.

A normalização de dados utilizada foi aquela descrita por (eq:norm3), na subseção 3.3.1. A justificativa para a escolha desse tipo de normalização de dados, além do aspecto relacionado ao degrau de consumo, é que ele é capaz de tornar comparáveis padrões cujo consumo mensal médio é muito distinto. Uma crítica a este sistema é que o mesmo depende do degrau de consumo, o que não pode ser detectado se a fraude for executada ao mesmo tempo que a conexão da UC à rede, como é o caso de ações ilegais implementadas ainda na construção de uma casa, por exemplo. Isso pode ser contornado ou ao menos mitigado se a distribuidora possuir um estudo amplo sobre as posse de equipamentos de seus clientes. Na literatura encontra-se exemplo da utilização desses estudos para o melhoramento de sistemas baseados em IA para identificação e detecção de PNT (RIBEIRO et al., 2011). Como essa informação sobre estudo de posse de equipamentos ou potência instalada não estava disponível nas bases de dados das empresas consultadas, decidiu-se por abandonar esses dados em um primeiro momento e focar nos padrões de degrau de consumo para projeto do sistema de reconhecimento de padrões.

5.3.2 Dados Categóricos

Em testes preliminares utilizando apenas históricos mensais de consumo, ficou evidente que uma opção para melhorar o desempenho do classificador seria aumentar a quantidade de informações de entrada com dados de outros tipos. Uma maneira de obter bons resultados com sistemas de reconhecimento de padrões é a incorporação de conhecimento sobre o problema (DUDA; HART; STORK, 2000). Na literatura que trata de PNT,

um fator citado recorrentemente é a correlação entre localização geográfica das UC e a frequência com que são encontradas PC (BASTOS; SOUZA; FERREIRA, 2009a; NAGI et al., 2010; QUEIROGA, 2005; DANTAS, 2006; PENIN, 2008; ORTEGA, 2008).

Os dados de entrada de sistemas de reconhecimento de padrões podem ser classificados em duas categorias: variáveis quantitativas e variáveis qualitativas. As variáveis quantitativas são aquelas representadas por valores numéricos, contínuos ou discretos. Como exemplo de variável quantitativa contínua podemos citar a energia medida em kWh em um medidor de energia elétrica. Um exemplo de variável quantitativa discreta é o número de cortes de energia realizados em uma unidade consumidora. As variáveis qualitativas, por sua vez, são dados pertencentes a categorias, logo não são descritos por valores numéricos. Quando existe relação de intensidade entre variáveis qualitativas de maneira que estas podem ser ordenadas, elas são chamadas de ordinais. Caso não haja essa relação, as variáveis são chamadas de nominais. Exemplos de variáveis qualitativas ordinais: nível de tensão (baixo, médio, alto), descrição da temperatura (frio, ameno, quente). Tipos de clientes (residencial, industrial, comercial, etc.) ou ramo de atividade (padaria, supermercado, posto de gasolina, etc.) são exemplos de variáveis qualitativas nominais (ORTEGA, 2008).

Os métodos estudados até o momento, incluindo o OPF, precisam de dados de entrada contendo valores numéricos para serem implementados, portanto é necessário fazer uma representação numérica de variáveis qualitativas para que estas possam ser utilizadas. O nome desse processo é codificação de dados (ORTEGA, 2008).

5.3.2.1 Codificação

A codificação dos dados consiste em encontrar uma representação numérica para dados qualitativos. No caso de variáveis ordinais, a codificação é um processo muito mais simples do que no caso de variáveis nominais. Uma maneira de se codificar variáveis ordinais é codificar as categorias sobre uma variável numérica e definir valores inteiros ordenados para cada categoria. Por exemplo: para os níveis de tensão baixa, média e alta poderíamos atribuir os valores -1", "0"e "1", respectivamente.

Um método simples de codificação de variáveis nominais encontrado na literatura em (MURPHY, 2012) e (ORTEGA, 2008) é chamado de *1-of-m encoding*, *dummy encoding* ou *one-hot encoding*. Supondo que haja m variáveis nominais, essas variáveis são representadas por vetores de m dimensões binárias, de maneira que para cada variável nominal só possa haver uma dimensão com valor "1"e para todas as outras o valor seja "0". Exemplo de codificação: cliente residencial $(1\ 0\ 0)$, cliente comercial $(0\ 1\ 0)$ e cliente industrial $(0\ 0\ 1)$.

Desta forma, no algoritmo de extração de características implementado, foi adicionada a capacidade de se utilizar dados categóricos nominais. O objetivo desse tipo de dados é passar a informação ao classificador sobre a correlação entre duas amostras pertencentes à mesma localidade.

5.4 Classificador OPF

O classificador OPF utilizado foi implementado utilizando os algoritmos apresentados no capítulo 4. Tendo em vista os resultados mostrados na literatura em (RAMOS et al., 2011), foi considerada a utilização do algoritmo de aprendizado. A figura 8 mostra em detalhes os blocos de extração de características e treinamento presentes na figura 7. As etapas que precedem o OPF, implementadas pelo algoritmo de extração de características,

são a codificação dos dados categóricos (DC) selecionados e também a normalização de dados pela média do período considerado. Assim, os dados de entrada do OPF são dados de consumo mensal e DC relativos à localização geográfica. Para considerar as variações sazonais de consumo (como por exemplo a mudança de hábitos dada pelas estações do ano, período de férias e feriados) o histórico de consumo mensal dos últimos 12 meses é utilizado.

Todos os dados rotulados, que são aqueles referentes às UC já investigadas das quais se conhece a classe a que pertencem, formarão os conjuntos de treinamento e aprendizado, Z_1 e Z_2 respectivamente, de tamanhos iguais. Os dados são alocados de maneira aleatória nesses dois conjuntos, excluindo combinações que contenham dados de apenas uma categoria. Esses dois conjuntos passam então por 10 iterações do algoritmo de aprendizado aplicado a PNT (algoritmo 2 da subseção 4.5.1). O número de iterações para o aprendizado foi determinado empiricamente após verificar-se que para a maioria dos casos analisados a curva de aprendizado raramente apresentava melhora após cerca de 10 iterações. A etapa de obtenção de protótipos é iniciada pela obtenção da MST A.3. O algoritmo selecionado para realização desta tarefa foi o algoritmo de Kruskal, pela sua simplicidade de implementação. Depois de obtida a MST, o procedimento de obtenção do conjunto S é como aquele descrito na seção 4.2, onde se determina que os protótipos são aqueles vértices de classes diferentes conectados por arestas da MST. Uma vez obtidos os protótipos, ocorre o treinamento do classificador (algoritmo 1 seção 4.3). O classificador treinado é testado classificando o conjunto Z_2. Em seguida, amostras são trocadas entre os conjuntos Z_1 e Z_2 de acordo com os critérios do algoritmo de aprendizado. Ao final das 10 iterações, o resultado é o classificador OPF treinado. Esse classificador é então aplicado no conjunto de dados não rotulados, Z_3, gerando uma lista de UC candidatas a inspeção.

O algoritmo de poda, tal qual é descrito na literatura, tem como objetivo a redução do tempo de processamento necessário para concluir a etapa de classificação do OPF (RAMOS et al., 2011). Isso é alcançando diminuindo o número de amostras de Z_1, que tem como consequência a redução do número de amostras passíveis de conquistar cada amostra a ser classificada, reduzindo assim o número de operações necessárias para realização da classificação. Como foi verificado em testes preliminares, apesar de cumprir o seu objetivo de acelerar de fato a etapa de classificação, o algoritmo de poda reduz o índice de acerto do classificador. Isso ocorre porque alguns ramos das árvores geradas pelo classificador OPF são consideradas irrelevantes durante classificação do conjunto de aprendizado Z_2 e são eliminadas pelo algoritmo de poda mas servem para classificar corretamente elementos do conjunto de treinamento Z_3. Como não houve maiores problemas com tempo de execução do OPF nos testes realizados, este algoritmo não foi utilizado.

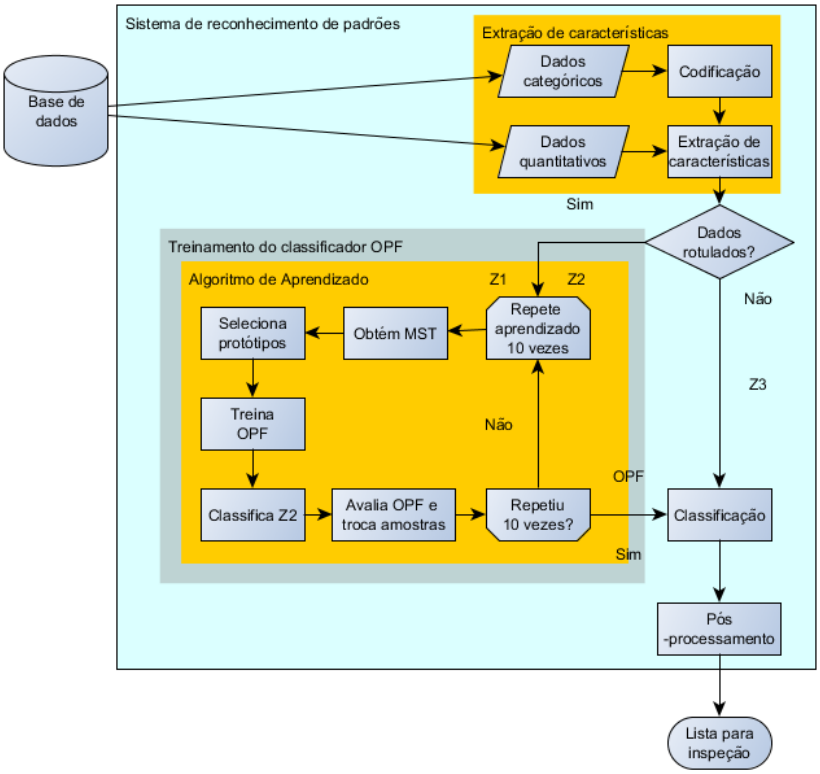

Figura 8: Detalhe do algoritmo OPF com aprendizado aplicado a PNT. Algoritmo de aprendizado avalia 10 vezes classificador gerado por Z_1 e troca amostras deste conjunto com Z_2, de aprendizado. Elementos de Z_3 não são rotulados, ou seja, são UC que não foram inspecionadas. Fonte: Próprio autor.

5.5 Pós-processamento

A lista de UC candidatas a inspeção gerada pelo OPF pode ainda receber uma última etapa de processamento que pode ter como objetivos a combinação destes resultados com outros métodos ou mesmo a seleção de alguns elementos da lista de acordo com alguma regra. Uma das maneiras de se melhorar o desempenho de classificadores é a sua combinação por meio de uma votação.

5.5.1 Votação

Com o objetivo de buscar uma alternativa que pudesse melhorar o desempenho do classificador OPF desenvolvido até então, decidiu-se pela utilização de um algoritmo de votação como etapa de pós-processamento do OPF. Este é baseado nos princípios dos sistemas *multi-net* comissionados citados em (ORTEGA, 2008). Sistemas *Multi-net* comissionados são aqueles que em que a tarefa de classificação é realizada por vários elementos que realizam a mesma tarefa de maneira independente, combinando os resultados de cada classificador em um resultado global utilizando um método de fusão. Segundo a mesma fonte, em muitas aplicações é possível obter resultados melhores por um comitê do que por classificadores individuais. Um dos métodos de fusão propostos é a votação. Para tal, treina-se vários classificadores a partir de bases de dados diferentes. Então o conjunto de teste é classificado pelos classificadores treinados e seu resultado é comparado. Por fim, a classificação final da amostra é aquela dada pela maior quantidade de classificadores. Uma versão deste método aplicado a curvas de carga foi apresentada em (TREVIZAN et al., 2014).

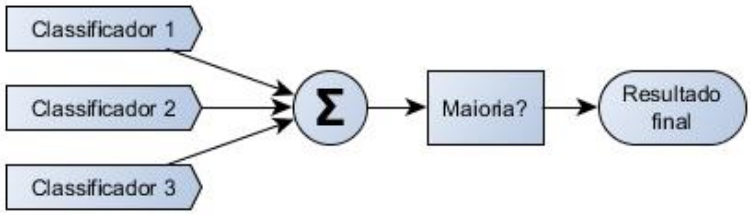

Figura 9: Esquema de um algoritmo de votação como aquele apresentado em (TREVIZAN et al., 2014). Fonte: Próprio autor.

Nos testes realizados não havia diferentes bases de dados para as amostras aplicadas nem quantidades diferentes de dados disponíveis para cada amostra. Como verificou-se um melhor desempenho do algoritmo de aprendizado do que aquele de votação, neste trabalho o resultado final do sistema de reconhecimento de padrões não passará por nenhuma etapa de pós-processamento.

5.6 Conclusão

Neste capítulo o sistema proposto para resolução do problema de identificação e localização de perdas comerciais é apresentado em detalhe. Os dados de entrada escolhidos para representar numericamente cada UC sob a forma de vetores de características são DC codificados juntamente com históricos de consumo mensal normalizados pelo consumo mensal médio do período considerado. Os DC são aqueles referentes à localização geográfica das UC, cuja importância já foi largamente ressaltada na literatura. Os dados de

consumo mensal normalizado tem como objetivo pôr em evidência a redução do consumo gerada pela inserção de PC em uma UC.

O processo para identificação dessas PC é implementado pelo processamento desses dados por um classificador OPF em uma versão que inclui uma etapa de aprendizado. Para melhores resultados, devem ser utilizados para projeto do classificador todos os dados rotulados disponíveis. O classificador OPF utilizado possui etapa de aprendizado. Por motivo de simplicidade de implementação, a etapa de obtenção da MST é implementada pelo algoritmo de Kruskal. A etapa de poda estudada em 4.5.2 não foi incluída pois sua utilização implica na redução do desempenho do classificador.

Uma vez treinado, o OPF é utilizado para classificar os elementos não rotulados da base de dados, gerando uma lista de UC candidatas a inspeção. Essa lista é então utilizada por equipes de inspeção que vão a campo confirmar a existência de PC nas UC apontadas.

6 ESTUDO DE CASO

6.1 Introdução

Uma vez idealizado, o sistema desenvolvido precisa ser testado. A exemplo do que se encontra na literatura que trata de sistemas de reconhecimento de padrões, o método de escolha para avaliação do desempenho do sistema proposto foi a sua aplicação utilizando dados de entrada de uma base de dados rotulada. Explicado de outra maneira, o que se pretende é a utilização de dados de UC já inspecionadas divididos em conjuntos para treinamento OPF e teste do seu desempenho. Idealmente, essa avaliação de desempenho seria feita com bases de dados provenientes de distribuidoras de energia elétrica. No entanto, tendo em vista que não foi possível a obtenção de uma base de dados rotulada, foi necessário desenvolver uma base de dados para teste do sistema.

Este capítulo trata dos testes realizados para verificação do desempenho do sistema proposto. Em um primeiro momento, a base de dados desenvolvida é apresentada. Em um segundo momento, são descritos os testes realizados. Em seguida os resultados são apresentados e analisados. Por fim, são descritas as conclusões a que se chegou a partir do presente estudo de caso.

6.2 Base de Dados

Embora tenha sido possível a obtenção de dados provenientes de um sistema real para testar o método proposto, não tem-se a disposição resultados de inspeções realizados nas unidades consumidoras do sistema real, necessários para a avaliação do desempenho dos algoritmos do OPF. Por conta dessa limitação, foi desenvolvida uma base de dados puramente fictícia para servir como base para que se possa verificar o funcionamento dos métodos apresentados até este momento. Essa base de dados, contendo dados de consumo mensal gerados a partir de dados na literatura é chamada neste trabalho de *sistema-teste* e tem como objetivo avaliar o funcionamento do método desenvolvido.

6.2.1 Sistema-Teste

O sistema-teste é uma base de dados que simula um alimentador de um sistema de distribuição, contendo todos os dados necessários para a aplicação das metodologias de classificação de consumidores suspeitos por meio do OPF, também sendo utilizado para avaliação de desempenho de outros métodos para localização e identificação de PNT desenvolvidos pelo grupo de pesquisa do qual o autor desta dissertação faz parte. Ele é composto por dados relativos à topologia do alimentador, barras onde se cada UC está conectada, curvas de cargas horárias em kW para cada cliente e históricos de consumo

mensais em kWh para cada unidade consumidora considerada. De posse dessas informações, pode-se alterar os dados de consumo mensal de algumas UC inserindo nelas comportamentos semelhantes àqueles gerados por PC. Como sabe-se de antemão em quais UC foram inseridas essas "perdas", tem-se uma base de dados rotulada sintetizada a partir de dados da literatura.

Os dados de topologia e potências das cargas do alimentador são baseados naqueles do sistema 123 barras da IEEE (*IEEE 123-Bus Test Feeder*, figura 10) (KERSTING, 2001). Esse sistema foi escolhido pois representa um alimentador de um sistema de distribuição em média tensão. No que diz respeito à distribuição do consumo mensal das UC, os dados foram retirados da bibliografia (JARDINI; CASOLARI, 1999). As variações mensais de consumo foram baseadas no comportamento sazonal de clientes de uma distribuidora de energia elétrica do Brasil.

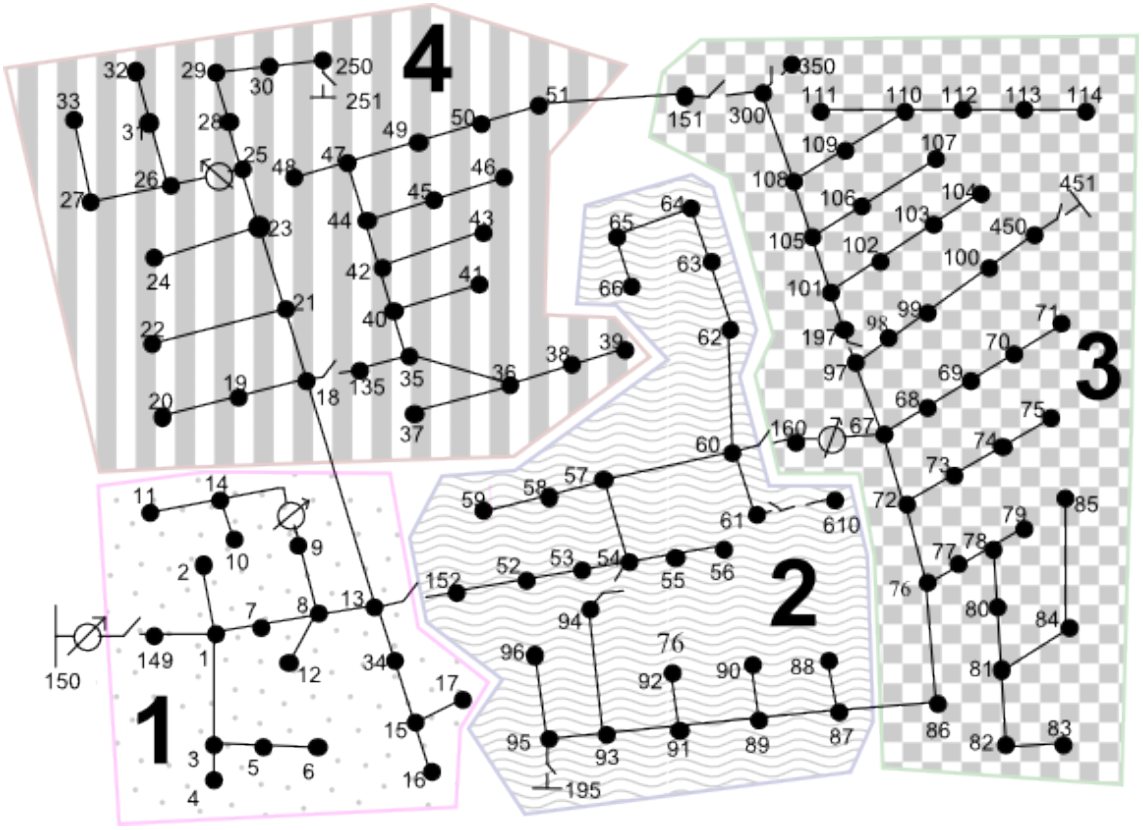

Figura 10: Sistema 123 barras da IEEE adaptado com divisão do sistema em 4 regiões. Adaptado de (KERSTING, 2001).

Tabela 4: Distribuição das barras carregadas por regiões.

Região	Barras
1	1, 2, 4, 5, 6, 7, 9, 10, 11, 12, 16, 17, 34
2	52, 53, 55, 56, 58, 59, 60, 62, 63, 64, 65, 66, 87, 88, 90, 92, 94, 95, 96
3	68, 69, 70, 71, 73, 74, 75, 76, 77, 79, 80, 82, 83, 84, 85, 86, 98, 99, 100, 102, 103, 104, 106, 107, 109, 111, 112, 113, 114
4	19, 20, 22, 24, 28, 29, 30, 31, 32, 33, 35, 37, 38, 39, 41, 42, 43, 45, 46, 47, 48, 49, 50, 51

A distribuição de clientes por faixa de consumo mensal é dada como na tabela 5. Na referência, a primeira faixa é de 0 a 50 kWh. Foi considerado que deveria haver um valor mínimo maior que zero, então arbitrou-se que este deveria ser 5 kWh. Por motivo de praticidade, dados para apenas 12 meses consecutivos foram criados.

Tabela 5: Distribuição de UC por consumo mensal.

Demanda [kWh]	**Número de UC [%]**
5 - 50	1,58
51 -100	8,40
101 - 150	22,28
151 - 200	27,03
201 - 250	17,95
251 - 300	11,57
301 - 400	7,92
401 - 500	0,00
501 - 1000	3,26

Fonte: (JARDINI; CASOLARI, 1999).

Foram utilizados dois modelos de perdas comerciais, baseados em (HUANG; LO; LU, 2013). O primeiro deles tem como objetivo simular casos como furto de energia e defeitos ou fraudes no medidor em que o medidor deixa de contar energia elétrica ou para completamente de funcionar (figura 11). Já o segundo simula uma caso em que a fraude ou defeito gera uma redução no valor da energia faturada pelo medidor ou quando uma derivação é inserida no ramal de entrada antes do medidor de forma que parte do circuito da UC é alimentado por energia que não é medida (figura 12).

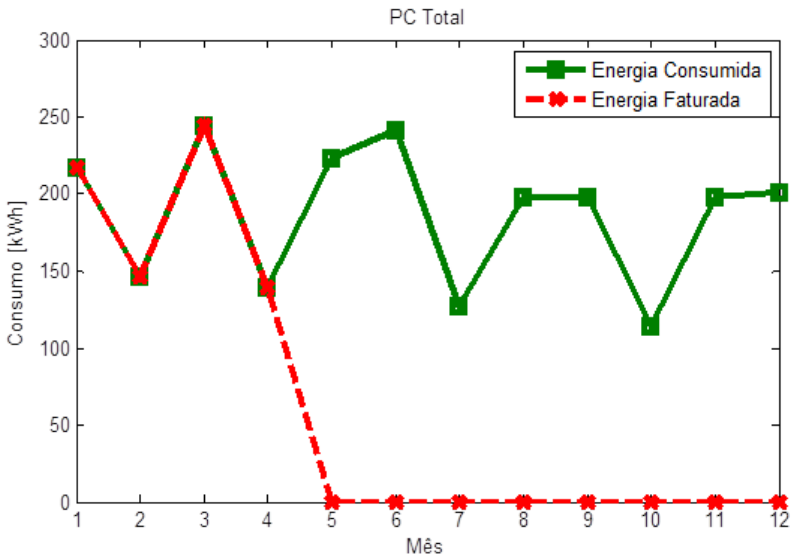

Figura 11: Modelo de perda comercial com redução total no consumo. Fonte: Próprio autor.

As potências médias das cargas do sistema de 123 barras foram mantidas. Desta maneira, cada barra deve apresentar consumo anual médio igual àquele do sistema de 123 barras. O número de clientes por barra será avaliado de acordo com a potência da barra, tal qual ela é descrita no sistema do IEEE. A potência média descrita no sistema do

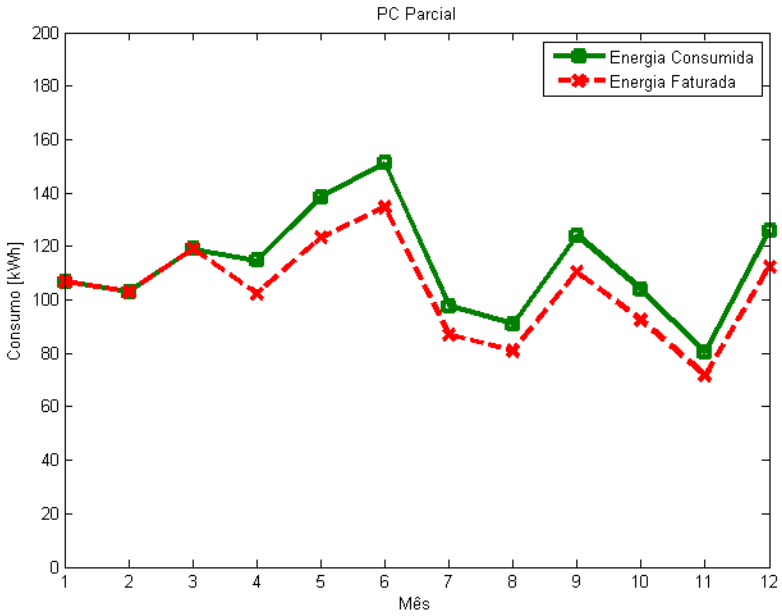

Figura 12: Modelo de perda comercial com redução parcial no consumo. Fonte: Próprio autor.

IEEE será igual à potência média anual dos clientes ligados a cada barra. Portanto, em alguns meses as cargas terão potências maiores do que aquelas descritas no 123 barras do IEEE e em outros meses a potência média será menor. No total foram geradas 12180 UC distribuídas entre as 85 barras do sistema que tem carga de acordo com a distribuição indicada na tabela 5.

De posse do número de clientes ligados a cada carga do sistema de 123 barras, é definida para cada cliente um consumo mensal. Assim, para cada cliente serão criados 12 consumos mensais. O comportamento sazonal do consumo mensal das unidades consumidoras é obtido a partir de dados extraídos do sistema de uma distribuidora. Isso foi realizado retirando da base de dados da empresa históricos mensais, separados a cada 12 meses. Cada curva de 12 meses foi então filtrada, eliminando aquelas que não possuíam consumo durante os 12 meses do ano, possivelmente causados pela desocupação da UC ou pela sua ligação à rede elétrica. Cada curva foi normalizada pela sua potência anual média, o que resultou em curvas de consumo anual adimensional e de média 1 para cada UC. A partir de todas essas curvas foi calculada uma curva média de todas as curvas médias e para cada mês foi obtido um valor de desvio padrão. Essas duas curvas são a representação da sazonalidade do consumo mensal extraída da base de dados da empresa (figura 13). A partir dessas curvas foram geradas as curvas de consumo mensal para cada cliente, de acordo com (15).

$$Consumo(mês_m) = n(1, \sigma_m) \cdot Consumo_{médio} \tag{15}$$

onde $Consumo(mês_m)$ é o consumo da UC no m-ésimo mês, $n(1, \sigma_m)$ é um número aleatório obtido a partir de uma distribuição normal com média 1 e desvio-padrão σ_m do m-ésimo mês e $Consumo_{médio}$ é o consumo médio anual da UC. Assim, será possível obter para qualquer mês de cada UC o seu histórico de consumo mensal. Então, por meio de um fluxo de carga, é possível saber o estado em qualquer ponto do Sistema de Referência no instante de tempo desejado. Uma segunda base de dados simula os dados obtidos pela concessionária. Essa base de dados contém os dados de fatura mensais em

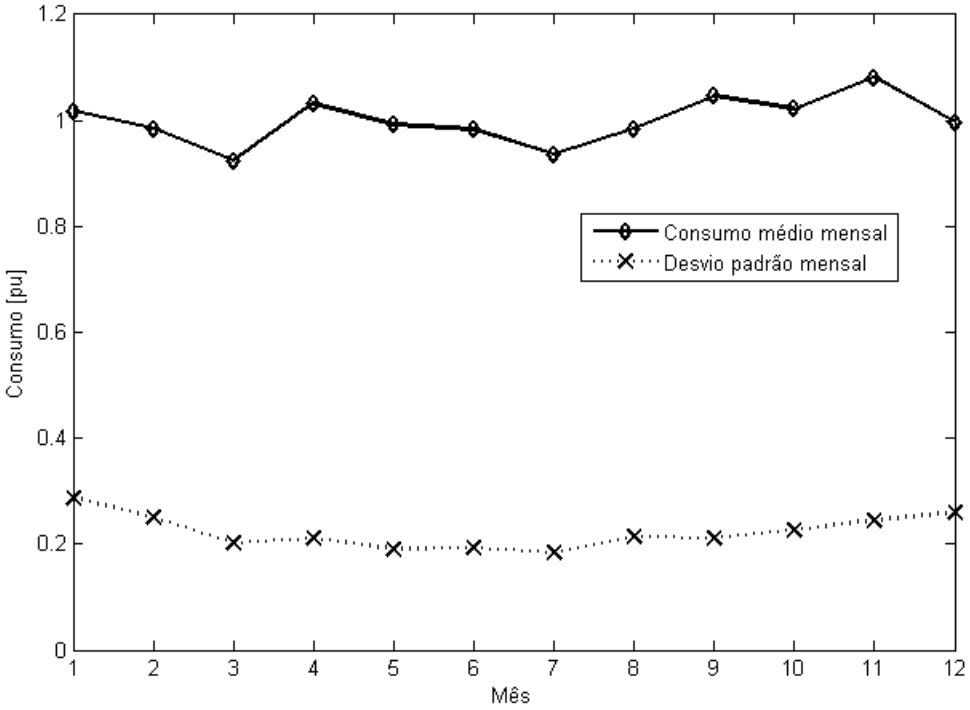

Figura 13: Perfil do consumo mensal médio dos clientes residenciais de um alimentador. Uma das curvas mostra a média mensal enquanto a outra mostra o desvio-padrão mensal. Fonte: Próprio autor.

kWh para cada consumidor em um período de 12 meses, como se fossem dados obtidos de leituras de medidores de energia elétrica. Esses dados que simulam leituras são chamados de casos . Foram criados 5 casos para testes, sendo que um deles, o caso-base, simula um sistema sem PC. Os demais casos, numerados de I a IV, simulam sistemas com PC. O objetivo desses casos foi avaliar o comportamento do método para diferentes modelos de PC (I e II) e também para a avaliação da hipótese de concentração de PC em certas regiões sobre o desempenho dos métodos de classificação. São eles:

Caso-base Não há PC. Serve de referência aos demais.

Caso I Perdas comerciais adicionadas aleatoriamente a aproximadamente 5% dos clientes. O modelo de PC utilizado é a diminuição da demanda a zero (simulando fraude ou defeito no medidor) a partir de um mês, escolhido também aleatoriamente, até o último mês do ano.

Caso II Versão modificada de I. Perdas comerciais adicionadas aleatoriamente a aproximadamente 5% dos clientes, mas com modelo diferente. O modelo de PC utilizado é a diminuição da demanda a um valor entre zero e o valor obtido do caso base a partir de um mês, escolhido também aleatoriamente, até o último mês do ano.

Caso III Perdas comerciais adicionadas aleatoriamente a aproximadamente 7% dos clientes, com base em dois modelos diferentes. Para 50% dos clientes onde há PC, o modelo de PC utilizado é a diminuição da demanda a um valor entre zero e o valor obtido do caso base a partir de um mês, escolhido também aleatoriamente, até o último mês do ano. Para os outros 50%, o modelo de PC utilizado é a diminuição da demanda

a zero a partir de um mês, escolhido também aleatoriamente, até o último mês do ano. Além disso, a distribuição das PCs entre os transformadores não é uniforme. Ela obedece uma distribuição que concentra as perdas nos primeiras barras onde há carga. Neste caso, 99% das UCs onde há PC estão nas 30 primeiras barras com carga (figura 14).

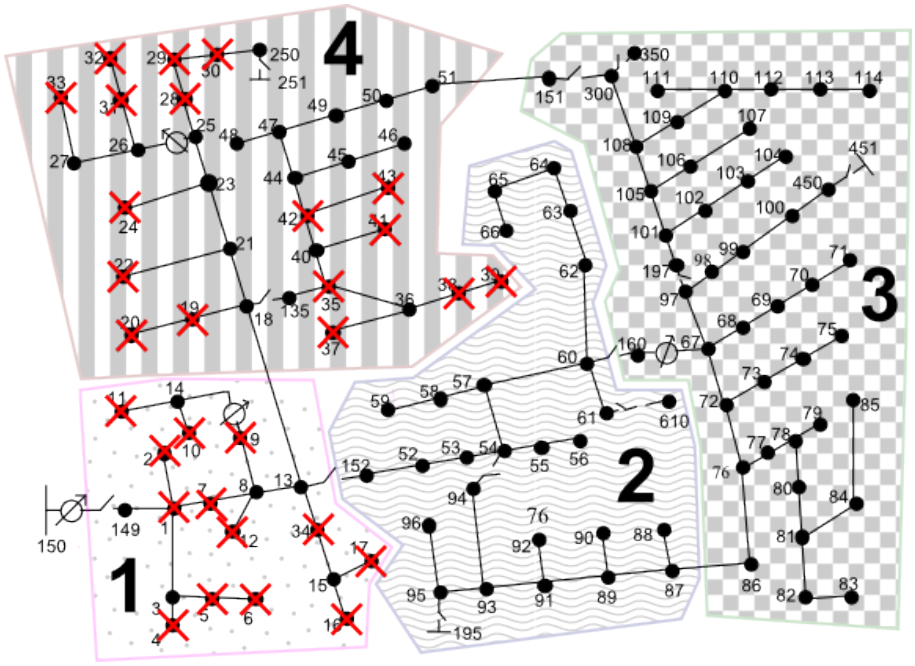

Figura 14: Localização de PC no caso III. Cruzes marcam locais com altos níveis de PC. Adaptado de (KERSTING, 2001).

Caso IV Versão modificada de III. Muda o desvio-padrão do modelo. Neste caso, 99% das UCs onde há PC estão nas 12 primeiras barras com carga (figura 15).

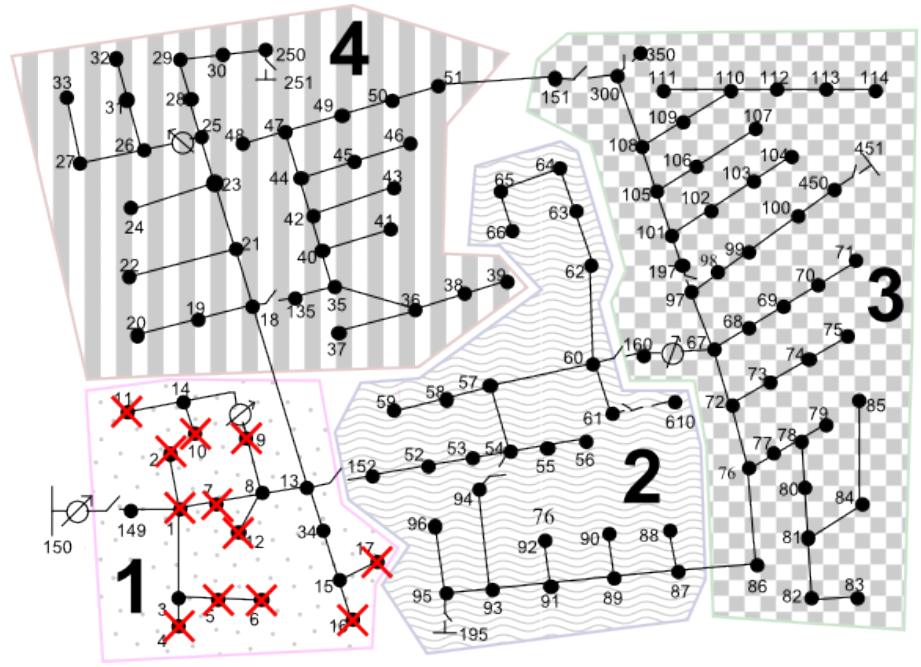

Figura 15: Localização de PC no caso IV. Cruzes marcam locais com altos níveis de PC. Adaptado de (KERSTING, 2001).

6.3 Testes Realizados

Testes para averiguação de desempenho foram realizados utilizando uma implementação do sistema de reconhecimento de padrões desenvolvido em linguagem MATLAB, realizada pelo autor. Diferentemente do que ocorreu em trabalhos de outros autores sobre o tema, a biblioteca LibOPF não foi utilizada. Nesta implementação foram utilizados o algoritmo OPF (algoritmo 1), o algoritmo de aprendizado (algoritmo 2) e o algoritmo de Kruskal para cálculo da MST.

O desempenho do algoritmo foi avaliado para uma série de fatores:

- comportamento do classificador para diferentes variações da base de dados;
- variação do tamanho do conjunto de treinamento;
- utilização de normalização de dados;
- utilização do algoritmo de aprendizado;
- desempenho para os indicadores Li, VPP, TCC, $SENS$ e F-score;
- utilização de DC;

Um estudo comparativo foi realizado com outros três métodos: ANN, SVM utilizando *kernel* linear (SVM-Linear) e SVM utilizando *kernel* RBF (SVM-RBF).

Os testes com o classificador OPF foram realizados alocando aleatoriamente os dados entre os conjuntos de dados de treinamento, aprendizado (onde há) e teste. Como o tamanho do conjunto de treinamento frente ao tamanho de todo o conjunto é um fator importante, foram testados valores considerados possíveis de serem obtidos na prática: 0,5% (60 amostras), 1% (121 amostras), 2% (243 amostras) e 5% (609 amostras) de dados nos conjuntos de treinamento ou treinamento e aprendizado. Para cada proporção

dos conjuntos, foram realizados 10 testes e o resultado final obtido foi uma média desses resultados.

6.3.1 Variações do OPF

Nos casos em que foi aplicado o algoritmo de aprendizado, este foi rodado por 10 iterações e cada conjunto utilizado no projeto, treinamento e aprendizado, ficou com 50% das amostras. O número de iterações foi determinado experimentalmente. Observou-se que os resultados para repetições do teste acima de 10 iterações exibiam indicadores de desempenho que se mantinham relativamente estáveis.

Neste caso foram testadas 10 variantes do método:

- Utilizando DC;
- Utilizando normalização dos dados;
- Combinações destes métodos.

Quando o OPF é testado sem o algoritmo de aprendizado, o programa segue um fluxo como mostrado na figura 16. Na etapa de extração de características pode haver ou não a utilização de dados categóricos, enquanto durante as etapas de treinamento e teste são 10 repetições.

Quando há a utilização do algoritmo de aprendizado, o fluxograma muda para o da figura 17. A diferença fica por conta pelas iterações do algoritmo de aprendizado. É importante notar que a parte mais demorada do aprendizado, o algoritmo para obtenção da MST, é rodado 100 vezes para cada caso testado.

6.3.2 OPF comparado com outros métodos

Um estudo comparativo foi realizado entre o OPF e os métodos ANN, SVM-Linear e SVM-RBF.

A ANN foi implementada por meio da *Neural Network Toolbox* do MATLAB. O projeto do classificador foi realizado pela ferramenta `nprtool` e a maior parte dos parâmetros sugeridos pela ferramenta foi adotada. A função utilizada para definição da rede foi `patternnet` e para treinamento foi `train`. Foi escolhida uma rede com 2 camadas ocultas e com configuração semelhante àquela utilizada em (RAMOS et al., 2009). Assim, o número de neurônios na entrada foi igual o número de atributos utilizados: 12 para o caso em que não foram considerados DC e 16 para o caso em que há DC. As camadas ocultas possuem cada uma número de neurônios igual a 1/4 do número de elementos do conjunto de treinamento. Na saída foram utilizados 2 neurônios. A função para classificação utilizada foi `net`. No caso em que se utilizou a etapa de validação-cruzada, 20% dos dados empregados no projeto do classificador foi alocado no conjunto de validação cruzada e os 80% restantes ficaram no conjunto de treinamento.

Os classificadores SVM foram treinados utilizando a função `svmtrain` e realizaram classificação por meio da função `svmclassify` ambas pertencentes ao *Statistics Toolbox* do MATLAB. A diferença entre eles foi na escolha do parâmetro que escolhe o *kernel*, `'kernel_function'`, que foi definido como `'linear'` no primeiro caso e `'rbf'` no segundo. Os parâmetros originais foram mantidos e nenhuma etapa de validação cruzada ou otimização de parâmetros do classificador SVM foi utilizada.

Os três classificadores foram utilizados em testes sobre o efeito da utilização ou não de DC e normalização de dados. Apenas o classificador ANN foi testado para etapa de validação cruzada, um análogo ao aprendizado no OPF.

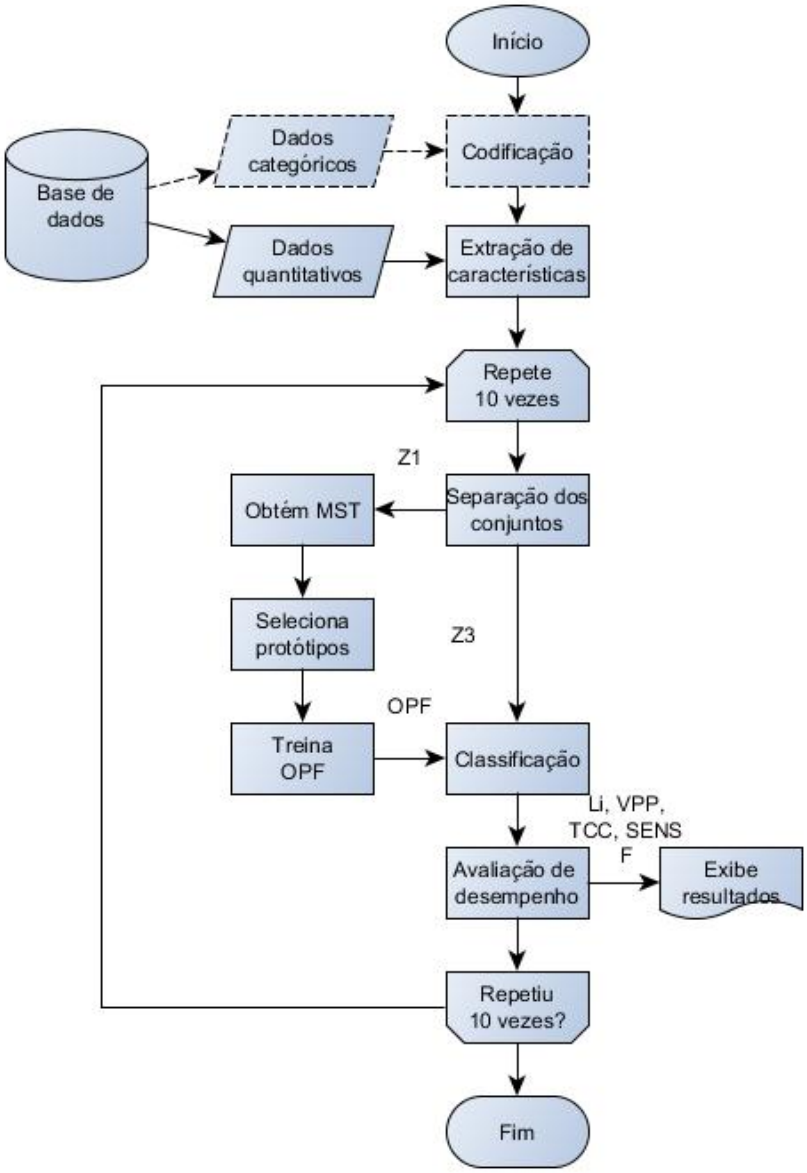

Figura 16: Etapas do teste do algoritmo OPF. Linhas tracejadas indicam etapas que são cumpridas apenas nos casos em que se consideram DC. Fonte: Próprio autor.

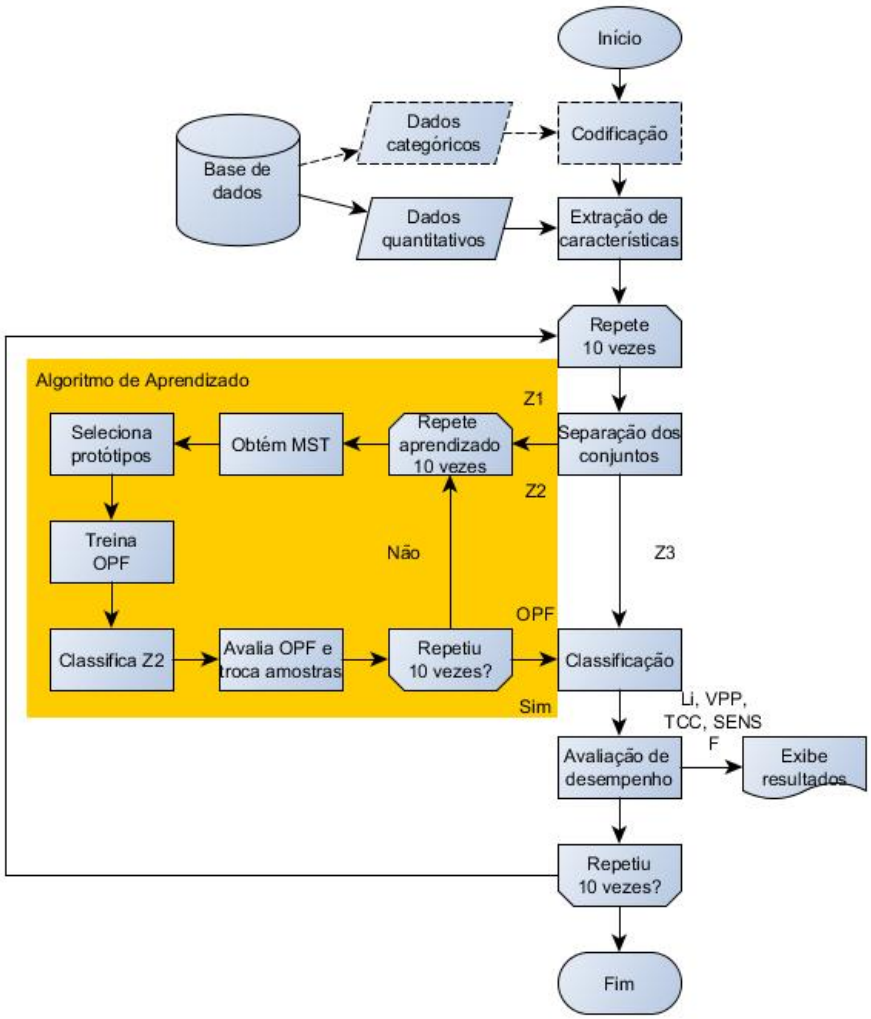

Figura 17: Etapas do teste do algoritmo OPF com aprendizado. Linhas tracejadas indicam etapas que são cumpridas apenas nos casos em que se consideram DC. Fonte: Próprio autor.

6.4 Resultados

6.4.1 Variantes OPF

O primeiro teste foi realizado para comparar os resultados da normalização dos dados, sem a utilização de dados categóricos ou de algoritmo de aprendizado, e seus resultados encontram-se na tabela 6. No segundo teste foi realizado para comparar os resultados da normalização dos dados, com a utilização de dados categóricos e sem a utilização de algoritmo de aprendizado, e seus resultados encontram-se na tabela 7. Na tabela 8 foi testado o efeito da normalização para os casos em que se utilizou o algoritmo de aprendizado, mas sem dados categóricos. O caso em que o efeito da normalização foi verificado com a utilização de dados categóricos e aprendizado tem seus resultados mostrados na tabela 9. Finalmente, o efeito da utilização algoritmo de aprendizado para o caso em que há consideração de DC e todas as características são normalizadas se encontra na tabela 10. Os melhores resultados obtidos em cada comparação estão em negrito e os melhores entre todos estão sublinhados.

Tabela 6: Comparação do uso ou não de normalização. Em negrito estão os melhores resultados. Não foram utilizados DC.

Caso	**Tr. [%]**	**Sem Normalização [%]**					**Com Normalização [%]**				
		Li	VPP	TCC	SENS	F	Li	VPP	TCC	SENS	F
I	0,5	66,98	**91,54**	**96,60**	34,19	**47,11**	**73,23**	59,59	91,31	**53,20**	41,01
	1	79,27	**92,75**	**97,76**	58,80	**71,54**	**80,61**	74,81	95,76	**63,83**	58,42
	2	82,95	**96,53**	**98,23**	66,03	**77,79**	**91,62**	73,11	97,45	**85,17**	77,19
	5	90,68	**98,07**	**99,03**	81,45	**88,94**	**95,50**	84,20	98,74	**91,90**	87,68
II	0,5	54,36	**43,39**	**93,21**	11,30	14,45	**58,26**	37,1	92,88	**19,90**	**23,28**
	1	55,65	26,46	92,28	15,05	16,39	**61,40**	**53,89**	**95,03**	**24,12**	**31,35**
	2	59,74	43,36	94,47	21,24	27,29	**64,79**	**46,12**	**94,48**	**31,88**	**37,03**
	5	63,36	41,33	94,46	28,90	33,86	**68,41**	**45,73**	**94,69**	**39,31**	**41,88**
III	0,5	57,44	**84,49**	**95,18**	15,05	24,76	**63,41**	67,87	93,51	**29,60**	**37,52**
	1	63,45	**82,10**	**95,63**	27,30	40,42	**69,38**	61,48	94,02	**41,71**	**46,12**
	2	70,06	**89,23**	**96,42**	40,45	54,48	**77,59**	70,93	95,96	**56,97**	**61,47**
	5	75,63	**91,43**	**97,08**	51,55	65,80	**80,75**	72,85	96,69	**62,85**	**67,43**
IV	0,5	**64,09**	**92,04**	**96,77**	28,33	**41,60**	62,40	44,92	89,49	**32,75**	26,14
	1	66,81	**90,49**	**96,92**	33,84	48,09	**71,04**	74,80	95,19	**44,60**	**50,73**
	2	72,21	**92,58**	**97,45**	44,59	**59,72**	**76,95**	70,96	96,50	**55,56**	59,17
	5	76,32	**94,41**	**97,81**	52,79	67,53	**81,11**	74,02	97,44	**63,24**	**68,19**

6.4.2 Comparativo OPF e outros métodos

Para reduzir a quantidade de dados exibida, nos testes comparativos optou-se pela utilização apenas do F-score. Este indicador foi selecionado pois o problema aqui tratado possui "classes desequilibradas"e os parâmetros mais importantes que deseja-se maximizar são VPP e $SENS$ simultaneamente, o que é contemplado pelo F-score. O primeiro teste do comparativo entre OPF, ANN, SVM-Linear e SVM-RBF foi realizado para comparar os resultados da normalização dos dados, sem a utilização de dados categóricos ou de algoritmo de aprendizado, e seus resultados encontram-se na tabela 11. No segundo teste foi realizado para comparar os resultados da normalização dos dados, com

Tabela 7: Comparação do uso ou não de normalização. Em negrito estão os melhores resultados. Há consideração de DC.

Caso	Tr. [%]	Sem Normalização [%]					Com Normalização [%]				
		Li	VPP	TCC	SENS	F	Li	VPP	TCC	SENS	F
I	0,5	70,30	86,76	**96,87**	40,87	52,66	**77,19**	**90,92**	96,10	**56,24**	**64,92**
	1	75,07	**94,38**	**97,45**	50,29	64,83	**84,30**	81,44	96,94	**70,31**	**72,74**
	2	84,44	**97,26**	**98,39**	69,00	**80,36**	**86,10**	68,85	96,20	**74,92**	68,33
	5	**90,84**	**98,93**	**99,07**	**81,72**	**89,42**	89,76	75,94	97,47	81,23	76,89
II	0,5	53,77	**47,42**	**94,29**	8,84	13,43	**57,79**	37,46	92,61	**19,20**	**22,96**
	1	58,38	36,33	**93,70**	19,24	22,96	**59,61**	**39,72**	93,55	**22,00**	**25,87**
	2	**60,70**	**36,33**	**93,92**	**23,89**	**27,83**	59,72	35,67	92,92	22,94	25,18
	5	**63,45**	40,00	94,25	**29,33**	33,40	63,43	**41,24**	**94,38**	29,13	**33,76**
III	0,5	63,19	**78,46**	**95,33**	27,09	39,32	**71,10**	74,49	94,07	**45,31**	**51,55**
	1	65,88	84,44	**95,92**	32,13	45,96	**73,95**	**93,65**	96,84	**48,23**	**61,94**
	2	70,06	**93,30**	**96,55**	40,32	55,85	**78,73**	66,74	95,35	**60,06**	**60,47**
	5	76,21	**90,45**	**97,10**	52,75	66,57	**82,67**	70,73	96,59	**67,04**	**68,52**
IV	0,5	67,43	77,19	96,70	35,40	47,85	**71,28**	**92,71**	**96,93**	**43,21**	**53,46**
	1	63,18	**82,87**	96,32	26,89	36,99	**78,82**	75,74	**96,78**	**59,15**	**62,54**
	2	72,82	85,66	97,26	46,07	59,32	**82,13**	**87,57**	**97,64**	**65,15**	**71,71**
	5	76,96	**94,98**	**97,88**	54,05	68,76	**85,85**	77,34	97,79	**72,79**	**74,21**

Tabela 8: Comparação do uso ou não de normalização. Em negrito estão os melhores resultados. O algoritmo de aprendizado é utilizado. DC não são utilizados.

Caso	Tr. [%]	Sem Normalização [%]					Com Normalização [%]				
		Li	VPP	TCC	SENS	F	Li	VPP	TCC	SENS	F
I	0,5	70,10	91,67	**96,85**	40,47	54,11	**79,87**	**91,70**	96,66	**61,28**	**68,87**
	1	79,36	**92,97**	**97,74**	59,01	70,91	**88,98**	82,06	97,25	**79,83**	**77,70**
	2	83,48	**97,66**	**98,32**	67,05	**79,16**	**91,70**	77,62	97,70	**85,06**	79,10
	5	88,68	**98,36**	**98,84**	**77,43**	86,54	**93,97**	86,85	98,73	**88,69**	**87,32**
II	0,5	54,39	34,73	93,53	11,02	13,91	**56,53**	**38,90**	**93,62**	**15,43**	**18,89**
	1	57,77	40,98	94,28	17,30	22,88	**61,69**	**51,65**	**94,79**	**25,01**	**32,46**
	2	60,61	34,57	93,84	23,79	27,41	**64,68**	**40,26**	**94,31**	**31,85**	**35,15**
	5	63,00	43,28	94,57	28,01	33,41	**67,67**	**50,38**	**95,00**	**37,38**	**42,36**
III	0,5	**66,79**	76,91	**95,68**	**34,35**	**45,51**	66,00	**91,38**	95,29	33,09	44,64
	1	65,07	**85,27**	**95,88**	30,47	44,02	**71,31**	80,10	95,65	**43,97**	**54,17**
	2	70,80	**84,74**	**96,36**	42,11	**55,43**	**73,82**	61,67	94,99	**50,02**	53,36
	5	76,20	**92,16**	**97,15**	52,67	**66,86**	**78,81**	74,90	96,60	**58,82**	65,52
IV	0,5	62,21	**83,70**	**96,14**	25,07	**35,91**	**66,88**	54,47	91,93	**39,47**	32,86
	1	66,79	**80,68**	**96,71**	34,03	46,70	**69,79**	70,78	95,34	**41,83**	**47,89**
	2	72,42	**87,92**	**97,35**	45,14	**58,49**	**75,86**	67,62	96,12	**53,70**	56,41
	5	77,94	**94,28**	**97,94**	56,03	**70,14**	**81,00**	77,05	97,49	**62,95**	68,64

Tabela 9: Resultado dos testes comparando uso ou não de normalização. Em negrito estão os melhores resultados. O algoritmo de aprendizado é utilizado. Há consideração de DC.

Caso	Tr. [%]	Sem Normalização [%]					Com Normalização [%]				
		Li	VPP	TCC	SENS	F	Li	VPP	TCC	SENS	F
I	0,5	70,10	**91,67**	**96,85**	40,47	54,11	**77,14**	90,87	96,06	**56,19**	**64,85**
	1	79,36	**92,97**	**97,74**	59,01	70,91	**84,28**	81,30	96,91	**70,30**	**72,62**
	2	83,48	**97,66**	**98,32**	67,05	**79,16**	**86,01**	68,55	96,15	**74,78**	68,07
	5	88,68	**98,36**	**98,84**	77,43	**86,54**	**89,58**	74,43	97,33	**81,00**	75,92
II	0,5	**57,13**	**42,67**	**93,97**	**16,30**	**20,35**	55,08	35,65	93,20	12,82	15,76
	1	56,63	30,58	93,01	16,32	19,76	**58,76**	**49,63**	**94,47**	**19,18**	**25,93**
	2	59,83	33,49	93,71	22,27	26,04	**60,53**	**35,87**	**94,10**	**23,33**	**27,74**
	5	64,32	39,23	94,08	**31,34**	34,21	**64,36**	**46,77**	**94,72**	30,72	**36,39**
III	0,5	60,25	84,84	95,36	20,82	31,74	**69,17**	**92,95**	**95,92**	**39,12**	**49,96**
	1	67,59	**86,71**	96,12	35,55	49,86	**72,77**	83,75	**96,16**	**46,50**	**57,76**
	2	69,63	**86,78**	**96,29**	39,68	53,68	**75,35**	65,37	95,43	**52,79**	**56,96**
	5	74,84	**89,22**	**96,93**	50,03	63,98	**80,50**	76,03	96,80	**62,20**	**68,07**
IV	0,5	63,67	**83,16**	**96,52**	27,71	39,84	**74,61**	65,77	95,47	**51,78**	**49,47**
	1	70,45	**88,33**	**97,13**	41,25	54,98	**74,75**	75,61	96,53	**50,89**	**57,34**
	2	70,68	**96,62**	**97,38**	41,44	56,86	**81,89**	73,12	97,01	**65,34**	**66,76**
	5	77,19	**91,36**	**97,82**	54,62	68,28	**86,26**	78,13	97,81	**73,61**	**74,92**

Tabela 10: Comparação do uso ou não de aprendizado. Em negrito estão os melhores resultados. Todos atributos são normalizados. Há consideração de DC.

Caso	Tr. [%]	Sem Aprendizado [%]					Com Aprendizado [%]				
		Li	VPP	TCC	SENS	F	Li	VPP	TCC	SENS	F
I	0,5	**80,38**	99,53	**98,09**	**60,76**	**74,22**	75,25	**99,95**	97,60	50,50	65,11
	1	**89,03**	**99,29**	**98,91**	**78,09**	**87,30**	84,17	98,29	98,42	68,39	79,34
	2	**91,16**	99,38	**99,12**	**82,35**	**89,89**	89,44	**99,48**	98,95	78,91	87,76
	5	93,11	99,41	99,31	86,24	92,30	**93,82**	**99,64**	**99,39**	**87,65**	**93,22**
II	0,5	56,60	39,03	93,61	15,58	19,07	**59,67**	**56,14**	**95,03**	**20,48**	**27,87**
	1	61,66	51,57	94,79	24,94	32,37	**63,86**	**63,70**	**95,29**	**29,02**	**37,96**
	2	**64,70**	41,39	94,38	31,80	35,46	64,27	**48,67**	**94,57**	**30,69**	**36,45**
	5	**67,67**	50,64	95,04	**37,35**	**42,49**	66,98	**53,11**	**95,25**	35,64	42,27
III	0,5	66,48	**99,19**	**96,30**	32,98	47,90	**68,24**	85,85	96,15	**36,89**	**50,99**
	1	71,85	93,55	96,74	43,88	59,16	**72,73**	**95,12**	**96,87**	**45,61**	**61,34**
	2	74,33	**94,75**	**97,02**	**48,83**	**63,93**	**73,35**	90,49	96,79	47,01	61,41
	5	78,83	91,34	97,39	57,98	70,82	**79,23**	**93,17**	**97,47**	**58,73**	**71,93**
IV	0,5	**69,53**	86,29	**97,04**	**39,41**	**51,52**	65,79	**86,85**	96,98	31,65	42,95
	1	70,80	94,19	97,29	41,81	56,79	**71,58**	**96,91**	**97,47**	**43,23**	**58,41**
	2	76,79	95,21	97,87	53,71	68,24	**77,20**	**95,84**	**97,94**	**54,50**	**69,29**
	5	81,36	93,15	98,18	62,95	74,98	**81,59**	**95,47**	**98,27**	**63,32**	**75,99**

a utilização de dados categóricos e sem a utilização de algoritmo de aprendizado, e seus resultados encontram-se na tabela 12. Na tabela 13 foram comparados os resultados dos classificadores para os casos em que foram considerados DC e os vetores de características completo de todos os elementos da base de dados foram normalizados pela sua média. O último teste realizado, cujos resultados se encontram na tabela 14, mostra os efeitos da utilização de normalização resultado normalização total ou parcial dos dados e também o efeito da utilização ou não de DC quando são aplicados os algoritmos de aprendizado (para o OPF) e validação cruzada (para ANN).

Para reduzir as dimensões das tabelhas, foram adotadas as seguintes abreviações:

S.N. sem normalização;

C.N. com normalização;

N.T. com normalização de todos atributos, inclusive dos DC;

Tabela 11: Comparação entre métodos com uso ou não de normalização. Em negrito estão os melhores resultados. Não há utilização de algoritmo de aprendizado ou validação cruzada. Não são considerados DC. O indicador utilizado é o F-score.

Caso	**Tr. [%]**	**OPF [%]**		**ANN [%]**		**SVM-Linear [%]**		**SVM-RBF[%]**	
		S.N.	C.N.	S.N.	C.N.	S.N.	C.N.	S.N.	C.N.
I	0,5	47,11	41,01	67,13	**76,56**	65,71	75,05	19,69	8,13
	1	71,54	58,42	83,47	75,06	76,77	**87,50**	55,04	20,11
	2	77,79	77,19	**90,86**	84,48	85,29	87,69	66,19	23,88
	5	88,94	87,68	**95,85**	91,96	89,68	90,97	74,27	37,13
II	0,5	14,45	23,28	24,55	25,13	**31,97**	31,66	7,17	0,03
	1	16,39	31,35	34,20	**38,70**	28,85	26,76	23,84	0,00
	2	27,29	37,03	38,75	**40,74**	29,96	28,58	26,50	0,07
	5	33,86	41,88	46,61	**49,55**	30,11	31,72	31,15	0,07
III	0,5	24,76	37,52	41,82	36,87	45,88	**55,21**	16,95	1,93
	1	40,42	46,12	**68,84**	60,06	60,38	59,92	31,34	9,18
	2	54,48	61,47	**72,45**	62,94	59,56	59,69	45,07	11,93
	5	65,80	67,43	**78,39**	76,19	53,83	57,06	60,83	16,65
IV	0,5	41,60	26,14	**57,12**	48,79	55,29	54,15	23,63	0,68
	1	48,09	50,73	**70,62**	55,54	57,12	50,48	21,23	9,08
	2	59,72	59,17	**71,54**	60,60	62,64	58,09	48,07	17,89
	5	67,53	68,19	**77,94**	76,59	62,10	59,17	62,25	26,23

Tabela 12: Comparação entre métodos com uso ou não de normalização. Em negrito estão os melhores resultados. Não há utilização de algoritmo de aprendizado ou validação cruzada. São considerados DC. O indicador utilizado é o F-score.

Caso	Tr. [%]	OPF [%]		ANN [%]		SVM-Linear [%]		SVM-RBF[%]	
		S.N.	C.N.	S.N.	C.N.	S.N.	C.N.	S.N.	C.N.
I	0,5	52,66	64,92	54,40	65,99	59,45	**79,24**	13,91	1,92
	1	64,83	72,74	73,36	69,67	80,23	**83,20**	8,39	4,60
	2	80,36	68,33	0,00	63,90	83,86	**90,16**	32,91	6,31
	5	89,42	76,89	0,00	0,00	90,16	**95,93**	55,57	15,85
II	0,5	13,43	22,96	20,32	33,74	20,31	**33,81**	3,65	0,00
	1	22,96	25,87	32,13	**34,52**	30,49	30,63	6,95	0,00
	2	27,83	25,18	11,16	**40,48**	31,30	33,52	15,09	0,00
	5	33,40	**33,76**	0,00	0,00	25,83	29,17	22,33	0,00
III	0,5	39,32	51,55	49,94	52,22	58,00	**61,39**	23,99	0,39
	1	45,96	61,94	**74,28**	69,42	67,83	72,11	31,79	5,42
	2	55,85	60,47	45,58	62,65	**74,49**	74,38	38,61	4,62
	5	66,57	**68,52**	0,00	0,00	66,24	67,62	59,13	14,71
IV	0,5	47,85	53,46	67,93	65,93	**72,49**	61,72	21,58	2,79
	1	36,99	62,54	**78,91**	71,53	75,16	77,34	42,81	9,83
	2	59,32	71,71	46,69	67,37	73,14	**78,34**	53,48	16,20
	5	68,76	74,21	0,00	0,00	**78,69**	77,46	64,94	26,92

Tabela 13: Comparação entre métodos com normalização de todas características com DC. Em negrito estão os melhores resultados. Não são utilizados algoritmo de aprendizado ou validação cruzada. São considerados DC. O indicador utilizado é o F-score.

Caso	Tr. [%]	OPF [%]	ANN [%]	SVM-Linear [%]	SVM-RBF [%]
I	0,5	74,22	73,23	**74,88**	1,91
	1	87,30	**89,20**	85,65	0,41
	2	89,89	89,64	**89,95**	6,23
	5	92,30	8,19	**96,03**	18,33
II	0,5	19,07	**35,21**	27,38	0,00
	1	32,37	35,35	**37,89**	0,00
	2	35,46	**44,98**	23,77	0,00
	5	**42,49**	5,83	30,51	0,00
III	0,5	47,90	**58,38**	56,96	1,90
	1	59,16	**77,28**	67,81	3,53
	2	63,93	78,31	**69,84**	6,91
	5	**70,82**	2,31	67,77	12,51
IV	0,5	51,52	**73,16**	70,48	5,99
	1	56,79	**78,47**	70,87	4,70
	2	68,24	**79,88**	77,88	15,12
	5	74,98	9,66	**77,74**	25,43

Tabela 14: Comparação entre métodos com uso ou não de normalização. Em negrito estão os melhores resultados. São utilizados algoritmo de aprendizado ou validação cruzada. São considerados DC. O indicador utilizado é o F-score.

Caso	Tr. [%]	OPF [%]					ANN [%]				
		Sem D.C.		Com D.C.			Sem D.C.		Com D.C.		
		S.N.	C.N.	S.N.	C. N.	N.T.	S.N.	C.N.	S.N.	C.N.	N.T.
I	0,5	54,11	68,87	54,11	64,85	65,11	36,14	71,68	13,47	61,74	**66,12**
	1	70,91	77,70	70,91	72,62	79,34	78,82	77,03	40,88	74,43	**88,04**
	2	79,16	79,10	79,16	68,07	87,76	76,87	78,37	0,00	73,86	**87,78**
	5	86,54	87,32	86,54	75,92	**93,22**	91,41	90,86	0,00	0,00	8,69
II	0,5	13,91	18,89	20,35	15,76	**27,87**	1,54	19,85	4,83	14,64	21,40
	1	22,88	32,46	19,76	25,93	**37,96**	10,15	31,03	1,95	21,31	26,65
	2	27,41	35,15	26,04	27,74	36,45	18,99	37,33	0,12	43,79	**46,96**
	5	33,41	42,36	34,21	36,39	42,27	37,24	**47,43**	0,00	0,00	5,56
III	0,5	45,51	44,64	31,74	49,96	**50,99**	18,74	36,99	28,89	44,36	45,88
	1	44,02	54,17	49,86	57,76	61,34	55,69	53,95	56,17	72,36	**70,97**
	2	55,43	53,36	53,68	56,96	61,41	67,62	**68,60**	27,95	65,79	68,33
	5	66,86	65,52	63,98	68,07	71,93	74,97	**77,52**	0,00	0,00	2,31
IV	0,5	35,91	32,86	39,84	49,47	42,95	32,78	44,91	57,13	55,59	**70,11**
	1	46,70	47,89	54,98	57,34	58,41	48,19	46,97	66,48	71,18	**77,06**
	2	58,49	56,41	56,86	66,76	69,29	60,62	67,49	36,83	73,87	**77,78**
	5	70,14	68,64	68,28	74,92	75,99	73,70	**77,12**	0,00	0,00	9,44

6.5 Análise dos resultados

6.5.1 OPF

O primeiro ponto a ser destacado é que para todos os casos analisados, o OPF em todas as suas variantes apresentou resultados médios muito superiores àqueles que seriam obtidos por um classificador aleatório, cujo VPP variaria entre 5%, nos casos I e II, e 7%, para os casos III e IV, o que demonstra que o método é eficaz. Se comparado aos resultados da indicação de clientes por especialistas das empresas distribuidoras de energia elétrica, conforme relatado na seção 3.1, o VPP dos métodos aqui desenvolvidos são igualmente superiores, o que justificaria a adoção do OPF em distribuidoras, somando-se à avaliação dos especialistas como parte do sistema de combate a PC das empresas.

Como era esperado, o desempenho dos métodos foi superior no caso I, cujo modelo de perdas adotado é aquele que deixa a existência de PC mais evidente, e inferior no caso II, quando a redução de consumo por conta de PC é em muitos casos imperceptível. Já nos casos III e IV o desempenho do OPF foi semelhante. O que também era esperado, mas que não ocorreu em todos os casos, foi o aumento progressivo do desempenho relativo ao aumento do tamanho do conjunto de treinamento. É razoável considerar que quanto mais se conhece sobre determinado conjunto de dados, o que é consequência da proporção de amostras que se possui no conjunto de treinamento, mais o classificador acerta. A exceção ocorre para alguns casos em que apenas um protótipo da classe "irregular"é identificado, na maioria dos casos o único desta classe no conjunto de treinamento, e possui muitos elementos próximos a si no conjunto de testes, o que leva a grande acerto (cuja consequência é um alto VPP) mas geralmente baixa abrangência na detecção de

elementos do conjunto de teste (dado por um baixo $SENS$).

Um efeito observado com relação à normalização é que, quando aplicada, ela costuma reduzir os índices VPP e TCC, mas aumentar os demais. Isso ocorre porque quando não há normalização das curvas de carga pela sua média, as diferentes potências médias anuais espalham os dados no espaço de características, de forma que quanto maior o consumo médio, mais afastado da origem do espaço se encontram as amostras de maior consumo médio. Logo, na etapa de treinamento há o aparecimento de mais "subpadrões", consequência do maior número de protótipos, para clientes normais e consumidores, o que de certa forma acaba resultado em *overfitting* do classificador. Desta forma, as UC com comportamento bastante semelhante aquele dos elementos da classe "irregular"contidos no conjunto de treinamento são classificados corretamente, mas são poucos. A partir dos resultados obtidos neste estudo de caso, pode-se concluir que ao se considerar a utilização de normalização deve-se esperar que haja um compromisso entre os indicadores VPP e $SENS$, sendo o primeiro maior quando não seu utiliza a normalização e o segundo quando os dados são normalizados.

A utilização de DC nos casos I e II, nos quais as PC foram distribuídas de maneira uniforme entre as barras, resultou em pontuações bastante próximas às dos testes em que os DC foram ignorados. Já para os casos III e IV, uma melhora pôde ser observada na maior parte dos casos. A partir desses resultados pode-se reafirmar a importância da utilização de DC no problema de localização e identificação de PC. No entanto, fica claro que o seu emprego é benéfico nos casos em que se sabe *a priori*, ou pelo menos suspeita-se, onde há altos níveis de concentração de PC e tais áreas são bem delimitadas.

O algoritmo de aprendizado apresentou diferentes comportamentos com relação para os diversos casos, não sendo possível afirmar que tenha havido uma tendência de melhora ou piora dos resultados da tabela 8 com relação àqueles da tabela 6. No entanto, quando utilizado em conjunto com a normalização de todos os atributos resultou na variação do OPF que teve melhores resultados.

6.5.2 Comparativo entre métodos

Nos testes em que foram considerados os DC com 5% de dados no conjunto de treinamento, o desempenho de ANN apresentou alguns resultados bastante inferiores aos demais métodos. Com exceção deste caso, pode-se dizer que a consideração de DC e a aplicação de normalização de dados melhoraram significativamente os resultados. De um modo geral, os melhores resultados foram obtidos com ANN, seguido de SVM-Linear e depois OPF.

Os resultados foram bastante diferentes daqueles descritos em (RAMOS et al., 2011, 2009). Nestes trabalhos, o OPF é o método que possui melhor desempenho em todos ou quase todos testes. Uma possível explicação para isso é que foram utilizadas diferentes bases de dados e implementações diferentes de cada método, já que as bibliotecas utilizadas não foram as mesmas. O SVM-RBF testado neste trabalho não foi implementado com etapa de otimização de parâmetros, por exemplo. Além disso, a base de dados utilizada neste trabalho foi sintetizada a partir de dados retirados de clientes residenciais, enquanto nos trabalhos supracitados os clientes escolhidos foram das classes comercial e industrial. Com relação aos indicadores escolhidos, em (RAMOS et al., 2009, 2011) temos avaliação de resultados utilizando Li, enquanto neste trabalho foi mostrado o *F-score* e o Li foi ocultado para não mostrar uma quantidade excessiva de dados. Com relação ao tempo de execução dos algoritmos, o OPF, principalmente quando utilizado o algoritmo de aprendizado, teve seu tempo de execução bastante superior aos métodos apresentados,

o que também difere bastante dos métodos encontrados na literatura, implementados em outras linguagens de programação. Foi observado que os testes realizados com as funções dos *toolboxes* do MATLAB operavam nos 4 núcleos do computador, enquanto o OPF implementado pelo autor utilizava apenas um deles.

6.6 Conclusão

Neste capítulo foram apresentados os testes realizados com o classificador OPF. A base de dados utilizada para os testes é proveniente de dados da literatura, sintetizada a partir de dados médios e desvios-padrão, utilizando como referência de topologia o sistema IEEE 123-Bus Test-Feeder. A partir de uma base de dados sintetizada sem a consideração de PC, 4 variantes desta foram criadas considerando a inserção de PC do tipo furto de energia, fraude no medidor e defeito no medidor.

O OPF foi testado com variações na sua implementação e utilização de dados de entrada. A consideração de DC foi considerada útil nos casos em que as regiões geográficas representadas pelos DC possuem altos níveis de PC, o que ocorre nos casos III e IV. A utilização de normalização dos dados resultou, de modo geral, em aumento nos indicadores Li, $SENS$ e *F-score*, reduzindo VPP e TCC. Como o problema tratado possui alto desequilíbrio na quantidade de elementos que cada classe, "regular"e "irregular", possui, o *F-score* foi considerado o indicador mais importante, logo a normalização dos dados pôde ser considerada benéfica ao classificador. O algoritmo de aprendizado considerado melhorou pouco os resultados quando utilizado isoladamente, e produziu a maior parte dos melhores resultados observados quando utilizado em associação com DC e normalização de todos os dados.

Em comparação aos métodos ANN, SVM-Linear e SVM-RBF implementados a partir das bibliotecas disponíveis no *software* MATLAB, o OPF obteve resultados significativamente superiores apenas ao último classificador citado. Diferentemente do que foi relatado em outros artigos, nos testes desenvolvidos neste trabalho, os métodos ANN e SVM-Linear obtiveram resultados melhores que o OPF em todas as suas variantes para a maioria dos testes realizados. O elevado tempo de execução do OPF torna evidente que a implementação em linguagem MATLAB utilizada precisa ser otimizada.

7 CONCLUSÃO

7.1 Conclusões gerais

As perdas no sistema elétrico brasileiro, especialmente as PC no sistema de distribuição, são um dos fatores que impedem maior modicidade tarifária, o que onera a sociedade brasileira, tornando o desenvolvimento de métodos para sua mitigação um tema relevante. Dentre as metodologias para este fim, este trabalho tratou daquelas cujo objetivo é a detecção e a identificação de PNT em sistemas de potência, tratando este problema pelo enfoque do reconhecimento de padrões. Analisando as obrigações impostas pela ANEEL com relação à obtenção e ao armazenamento de dados que poderiam ser utilizados em conjunto com esse tipo de metodologias, pode-se concluir que este enfoque poderia ser utilizado por um grande número de distribuidoras brasileiras, contribuindo para uma melhoria nos índices nacionais de PNT.

Dentre as metodologias pesquisadas durante a revisão de literatura, notou-se que a maioria delas utilizava classificadores supervisionados. O tipo de classificação realizada pela maioria era binária, isto é, rotulando os clientes analisados como suspeitos ou normais. Dentre os métodos utilizados pelos grupos de pesquisa com maior produção nos últimos anos aqueles que pareceram mais promissores foram os classificadores SVM, RNA e OPF. Dentre esses, o último foi o escolhido para o desenvolvimento deste trabalho.

A metodologia baseada no classificador OPF tal qual apresentada neste trabalho, considerando todas as premissas adotadas, indica que a sua utilização por empresas distribuidoras de energia elétrica pode contribuir para a redução da grande quantidade de PC no sistema elétrico brasileiro. Os experimentos realizados indicaram que as técnicas de normalização de dados e utilização de dados categóricos propostas neste trabalho podem aumentar o desempenho do OPF, principalmente quando utilizados em conjunto com métodos como o algoritmo de aprendizado.

A implementação em Matlab do OPF realizada pelo autor foi comparada com os classificadores RNA e SVM utilizando *kernel* Linear e RBF obtidos a partir de *toolboxes* do mesmo programa. Os resultados obtidos neste trabalho colocaram o RNA como melhor classificador para a maioria dos casos analisados, diferentemente do que foi encontrado em trabalhos que comparam os mesmos classificadores. Conclui-se que casos de estudo e, principalmente, as diferentes implementações são fatores cruciais na comparação entre métodos.

7.2 Trabalhos Futuros

Durante o desenvolvimento deste trabalho, ficou claro que alguns pontos poderiam ser melhorados. Para reduzir o tempo de processamento necessário no treinamento do OPF,

poderia-se utilizar a versão do OPF eficiente para grandes bases de dados (PAPA et al., 2012), além da utilização de algoritmo de Prim para obtenção da MST. Uso de etapas de melhoramento de desempenho para os algoritmos de Kruskal e Prim, ou mesmo outros algoritmos como de Boruvka poderiam ser igualmente testados. Outro ponto a explorar seria a utilização de paralelização de rotinas para acelerar execução de algoritmos, como a de classificação, que não é sequencial. Um melhoramento possível ao algoritmo OPF que não pôde ser implementado por falta de tempo é a modificação do índice de avaliação de desempenho do classificador no algoritmo do OPF para um mais adequado para a aplicação como o *F-score*.

Associação do OPF ou outros métodos baseados em conhecimento a métodos baseados na natureza física do problema como fluxo de carga ou estimador de estados está sendo desenvolvido pelo grupo de pesquisa, mas os resultados preliminares[1] [2] não foram incluídos por saírem do escopo deste trabalho. Em um destes trabalhos também foi utilizado um método de classificação não-supervisionado, aplicável a casos em que não há registros de inspeções em UC, não foi incluído neste trabalho pois também fugiu do escopo, mas tem sido desenvolvido pelo grupo de pesquisa.

A aplicação das metodologias a sistemas de distribuidoras e juntamente com a sua verificação metodologia é certamente o próximo passo lógico na continuidade desta pesquisa. Além disso, utilizando dados de bases de distribuidoras como estudos de posse de equipamentos como maneira de enriquecer a base de dados utilizada. Para melhor comparação das metodologias desenvolvidas com outras presentes na literatura seria ideal a utilização das mesmas bibliotecas e mesmas bases de dados.

[1] ROSSONI, A., et. al. Hybrid formulation for technical and non-technical losses estimation and identification in distribution networks: application in a Brazilian power system. In: CIRED 2015. **Proceedings...** [S.l:s.n]. Submetido.

[2] TREVIZAN, R. D., et. al. Identification of non-technical losses in typical distribution systems using OPF and state estimation. In: POWERTECH 2015. **Proceedings...** [S.l:s.n]. Submetido.

REFERÊNCIAS

AGÊNCIA NACIONAL DE ENERGIA ELÉTRICA (Brasil). **Resolução normativa 414**. 1.ed. [S.l.: s.n.], 2010. Disponível em: <http://www.aneel.gov.br/cedoc/ren2010414.pdf>. Acesso em 22 out. 2014.

AGÊNCIA NACIONAL DE ENERGIA ELÉTRICA (Brasil). **Procedimentos de distribuição de energia elétrica no sistema elétrico nacional**: módulo 7 - cálculo de perdas na distribuição. 4.ed. [S.l.: s.n.], 2013a. Disponível em <http://www.aneel.gov.br/arquivos/PDF/Modulo7_3-Final.pdf>. Acesso em 10 nov. 2014.

AGÊNCIA NACIONAL DE ENERGIA ELÉTRICA (Brasil). **Perdas de Energia**. Disponível em <http://www.aneel.gov.br/area.cfm?idArea=801>. Acesso em 14 nov. 2014.

AGUERO, J. Improving the efficiency of power distribution systems through technical and non-technical losses reduction. In: IEEE PES TRANSMISSION AND DISTRIBUTION CONFERENCE AND EXPOSITION (T&D), 2012, Orlando. **Proceedings...** [S.l.: s.n.], 2012. p.1–8.

ANGELOS, E. et al. Detection and identification of abnormalities in customer consumptions in power distribution systems. **IEEE Transactions on Power Delivery**, [S.l.], v.26, n.4, p.2436–2442, Oct 2011.

BALABANIAN, N.; BICKART, T. A. **Linear network theory**: analysis, properties, design and synthesis. Chesterland, Ohio, Estados Unidos: Matrix Pub, 1981.

BASTOS, P. R. F. d. M.; SOUZA, B. A. d.; FERREIRA, N. Uso de rede bayesiana na identificação das perdas não técnicas. In: ENCUENTRO REGIONAL IBEROAMERICANO DE CIGRÉ, 13., 2009, Puerto Iguazú. **Anales...** [S.l.: s.n.], 2009. p.1–9.

BASTOS, P. R.; SOUZA, B. A.; FERREIRA, N. Diagnosis of nontechnical energy losses using Bayesian Networks. In: INTERNATIONAL CONFERENCE AND EXHIBITION ON ELECTRICITY DISTRIBUTION, 20., 2009, Prague. **Proceedings...** [S.l.: s.n.], 2009. p.1–4.

CHANG, C.-C.; LIN, C.-J. **LIBSVM**: a library for support vector machines, software. 2001.

CRUZ, R.; QUINTERO, C.; PÉREZ, F. Detecting Non-Technical Losses in Radial Distribution System Transformation Point through the Real Time State Estimation Method. In: IEEE/PES TRANSMISSION & DISTRIBUTION CONFERENCE AND EXPOSITION: LATIN AMERICA, 2006, Caracas. **Proceedings...** [S.l.: s.n.], 2006. p.1–5.

DANTAS, P. R. P. **Avaliação de perdas de energia elétrica não técnicas, metodologia aplicada ao município de Salvador-Ba**. 2006. 96p. Dissertação (Mestrado em Regulação da Indústria de Energia) — Universidade de Salvador, Salvador, 2006.

DEPURU, S. S. S. R. et al. A hybrid neural network model and encoding technique for enhanced classification of energy consumption data. In: IEEE POWER AND ENERGY SOCIETY GENERAL MEETING, 2011, Detroit. **Proceedings...** [S.l.: s.n.], 2011. p.1–8.

DEPURU, S. S. S. R.; WANG, L.; DEVABHAKTUNI, V. Support vector machine based data classification for detection of electricity theft. In: IEEE/PES POWER SYSTEMS CONFERENCE AND EXPOSITION (PSCE), 2011, Phoenix. **Proceedings...** [S.l.: s.n.], 2011. p.1–8.

DRESCH, R. d. F. V. **Análise do efeito da modelagem da carga nas estimativas de perdas elétricas em sistemas de distribuição**. 2014. 70p. Dissertaçao (Mestrado em Engenharia Elétrica) — Universidade Federal do Rio Grande do Sul, Porto Alegre, 2014.

DUDA, R. O.; HART, P. E.; STORK, D. G. **Pattern classification**. 2nd.ed. [S.l.]: John Wiley & Sons, 2000.

EMPRESA DE PESQUISA ENERGÉTICA. **Plano Nacional de Energia 2030**: análise retrospectiva. Brasília: [s.n.], 2007. v.1. Disponível em <http://www.epe.gov.br/PNE/20080512_1.pdf>. Acesso em 13 nov. 2014.

EMPRESA DE PESQUISA ENERGÉTICA. **Plano Decenal de Expansão de Energia 2022**. Brasília: [s.n.], 2013. Disponível em <http://www.epe.gov.br/PDEE/20140124_1.pdf>. Acesso em 12 nov. 2014.

FERREIRA, J. B. **Mineração de dados na retenção de clientes em telefonia celular**. 2005. 93p. Dissertação (Mestrado em Engenharia Elétrica) — Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2005.

HAYKIN, S. **Neural networks**: a comprehensive foundation. 2nd.ed. Upper Saddle River, NJ, USA: Prentice Hall PTR, 1998.

HERNANDES JR, L. J. et al. Processo não invasivo de baixo custo para otimizaão da rotina de inspeão na detecão de furto de energia elétrica. **Revista Pesquisa e Desenvolvimento da ANEEL - P&D**, [S.l.], n.5, p.47–51, Ago 2013.

HUANG, S.-C.; LO, Y.-L.; LU, C.-N. Non-technical loss detection using state estimation and analysis of variance. **IEEE Transactions on Power Systems**, [S.l.], v.28, n.3, p.2959–2966, Aug 2013.

IGLESIAS, J. Follow-up and Preventive Control of Non-Technical Losses of Energy in C.A. Electricidad de Valencia. In: IEEE/PES TRANSMISSION DISTRIBUTION

CONFERENCE AND EXPOSITION: LATIN AMERICA, 2006, Caracas. **Proceedings...** [S.l.: s.n.], 2006. p.1–5.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **IPCA de outubro fica em 0,42%**. Disponível em <http://cod.ibge.gov.br/3JVPO>. Acesso em 16 nov. 2014.

JARDINI, J. A.; CASOLARI, R. P. **Curvas de carga de consumidores e aplicações na engenharia da distribuição**. São Paulo, Brasil: [s.n.], 1999. Disponível em: <http://stoa.usp.br/jajardini/files/1855/10320/1999-Jardini-Livro-Curva-de-carga.zip>. Acesso em 9 dez. 2013.

KERSTING, W. Radial distribution test feeders. In: IEEE POWER ENGINEERING SOCIETY WINTER MEETING, 2001, Columbus. **Proceedings...** [S.l.: s.n.], 2001. p.908–912, v.2.

KREYSZIG, E. **Advanced engineering mathematics**. 7th.ed. [S.l.]: Wiley, 1997.

LEÓN, C. et al. Variability and trend-based generalized rule induction model to NTL detection in power companies. **IEEE Transactions on Power Systems**, [S.l.], v.26, n.4, p.1798–1807, 2011.

MÉFFE, A. **Metodologia para cálculo de perdas técnicas por segmento do sistema de distribuição**. 2001. 139p. Dissertação (Mestrado em Engenharia Elétrica) — Universidade de São Paulo, São Paulo, 2001.

MÉFFE, A. **Cálculo de perdas técnicas em sistemas de distribuição** : modelos adequáveis às características do sistema e à disponibilidade de informações. 2007. 157p. Tese (Doutorado em Engenharia Elétrica) — Universidade de São Paulo, São Paulo, 2007.

MÉFFE, A.; de Oliveira, C. C. B. Technical loss calculation by distribution system segment with corrections from measurements. In: INTERNATIONAL CONFERENCE AND EXHIBITION ON ELECTRICITY DISTRIBUTION, 20., 2009, Prague. **Proceedings...** [S.l.: s.n.], 2009. p.1–4.

MONEDERO, I. et al. Detection of frauds and other non-technical losses in a power utility using Pearson coefficient, Bayesian networks and decision trees. **International Journal of Electrical Power & Energy Systems**, [S.l.], v.34, n.1, p.90–98, 2012.

MURPHY, K. P. **Machine learning**: a probabilistic perspective. [S.l.]: MIT press, 2012.

NAGI, J. et al. Detection of abnormalities and electricity theft using genetic support vector machines. In: IEEE REGION 10 CONFERENCE TENCON, 2008, Hyderabad. **Proceedings...** [S.l.: s.n.], 2008a. p.1–6.

NAGI, J. et al. Nontechnical loss detection for metered customers in power utility using support vector machines. **IEEE Transactions on Power Delivery**, [S.l.], v.25, n.2, p.1162–1171, 2010.

NAGI, J. et al. Improving SVM-based nontechnical loss detection in power utility using the fuzzy inference system. **IEEE Transactions on Power Delivery**, [S.l.], v.26, n.2, p.1284–1285, 2011.

NIZAR, A. H.; DONG, Z.; WANG, Y. Power utility nontechnical loss analysis with extreme learning machine method. **IEEE Transactions on Power Systems**, [S.l.], v.23, n.3, p.946–955, Aug 2008.

NIZAR, A. H.; DONG, Z. Y. Identification and detection of electricity customer behaviour irregularities. In: IEEE/PES POWER SYSTEMS CONFERENCE AND EXPOSITION, 2009, Seattle. **Proceedings...** [S.l.: s.n.], 2009. p.1–10.

NIZAR, A. H. et al. A data mining based NTL analysis method. In: IEEE POWER ENGINEERING SOCIETY GENERAL MEETING, 2007, Tampa. **Proceedings...** [S.l.: s.n.], 2007. p.1–8.

NORTE ENERGIA. **Conheça a UHE Belo Monte**. [S.l.: s.n.], 2011. Disponível em <http://www.blogbelomonte.com.br/wp-content/uploads/2011/11/folheto.pdf>. Acesso em 14 nov. 2014.

ORTEGA, G. V. C. **Redes neurais na identificaçao de perdas comerciais do setor elétrico**. 2008. 184p. Dissertaçao de Mestrado, Departamento de Engenharia Elétrica — Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio), Rio de Janeiro, 2008.

PAPA, J. P. et al. Design of robust pattern classifiers based on optimum-path forests. In: MATHEMATICAL MORPHOLOGY AND ITS APPLICATIONS TO IMAGE AND SIGNAL PROCESSING, 2007, Rio de Janeiro. **Proceedings...** [S.l.: s.n.], 2007. p.337–348.

PAPA, J. P. et al. Efficient supervised optimum-path forest classification for large datasets. **Pattern Recognition**, [S.l.], v.45, n.1, p.512–520, 2012.

PAPA, J. P.; FALCÃO, A. X.; SUZUKI, C. T. Supervised pattern classification based on optimum-path forest. **International Journal of Imaging Systems and Technology**, [S.l.], v.19, n.2, p.120–131, 2009.

PENIN, C. A. d. S. **Combate, prevenção e otimização das perdas comerciais de energia elétrica**. 2008. 214p. Tese (Doutorado em Engenharia Elétrica) — Universidade de São Paulo, São Paulo, 2008.

PYLE, D. **Data preparation for data mining**. [S.l.]: Morgan Kaufmann, 1999. v.1.

QUEIROGA, R. M. **Uso de técnicas de data mining para detecção de fraudes em energia elétrica**. 2005. 146p. Dissertaçao (Mestrado do Programa de Pós-Graduaçao em Informática) — Universidade Federal do Espírito Santo, Vitória, 2005.

RAMOS, C. C. O. et al. Fast non-technical losses identification through optimum-path forest. In: INTERNATIONAL CONFERENCE ON INTELLIGENT SYSTEM APPLICATIONS TO POWER SYSTEMS, 15., 2009, Curitiba. **Proceedings...** [S.l.: s.n.], 2009. p.1–5.

RAMOS, C. C. O. et al. A new approach for nontechnical losses detection based on optimum-path forest. **IEEE Transactions on Power Systems**, [S.l.], v.26, n.1, p.181–189, Feb 2011.

RAMOS, C. C. O. et al. What is the importance of selecting features for non-technical losses identification? In: IEEE INTERNATIONAL SYMPOSIUM ON CIRCUITS AND SYSTEMS (ISCAS), 2011, Rio de Janeiro. **Proceedings...** [S.l.: s.n.], 2011. p.1045–1048.

RAMOS, C. C. O. et al. New insights on nontechnical losses characterization through evolutionary-based feature selection. **IEEE Transactions on Power Delivery**, [S.l.], v.27, n.1, p.140–146, 2012.

RIBEIRO, F. B. et al. Prospecção de fraudes e identificação de ações de combate a perdas comerciais. **Revista Pesquisa e Desenvolvimento da ANEEL - P&D**, [S.l.], n.4, p.79–82, Ago 2011.

ROSSONI, A. et al. Load models effects on distribution system losses estimation: a numerical study. In: IEEE POWER AND ENERGY SOCIETY GENERAL MEETING (PES), 2013, Vancouver. **Proceedings...** [S.l.: s.n.], 2013. p.1–5.

SMITH, T. B. Electricity theft: a comparative analysis. **Energy Policy**, [S.l.], v.32, n.18, p.2067–2076, 2004.

SURIYAMONGKOL, D. **Non-technical losses in electrical power systems**. 2002. 90p. Dissertação (Mestrado em Ciência da Computação) — Ohio University, Athens, OH, 2002.

TREVIZAN, R. D. et al. Detecção e identificação de perdas comerciais usando curvas de carga e classificador floresta de caminhos ótimos. In: CONGRESO INTERNACIONAL DE DISTRIBUCIÓN ELÉCTRICA, 2014, Buenos Aires. **Anales...** [S.l.: s.n.], 2014. p.1–5.

U. S. ENERGY INFORMATION ADMINISTRATION. **International Energy Statistics**. Disponível em <http://www.eia.gov/cfapps/ipdbproject/IEDIndex3.cfm>. Acesso em 14 nov. 2014.

VIDINICH, R.; NERY, G. A. L. Pesquisa edesenvolvimento contra o furto de energia. **Revista Pesquisa e Desenvolvimento da ANEEL - P&D**, [S.l.], n.3, p.15, Jun 2009.

WEST, D. B. **Introduction to graph theory**. 2nd.ed. [S.l.]: Prentice Hall, 2001.

WILSON, R. J. **Introduction to graph theory**. 4th.ed. [S.l.]: Prentice Hall, 1996.

ANEXO A ELEMENTOS DE TEORIA DE GRAFOS

A.1 Definições

Um **grafo linear**, ou simplesmente **grafo**, é um conjunto de vértices, também chamados de nós, e arestas, chamadas de ramos (ou arcos). Cada ramo é conectado a dois nós distintos, não existindo ramos sem conexão. O oposto é possível: podem existir nós não conectados a ramo algum, chamados de nós isolados. Os ramos podem possuir orientação, o que nesse caso configura um grafo orientado (dígrafo ou grafo direcionado) (BALABANIAN; BICKART, 1981). De maneira mais formal, podemos definir um grafo G como sendo composto de dois conjuntos finitos: um conjunto V de pontos (vértices ou nós) e um conjunto E de linhas, chamadas ramos ou arestas, de tal maneira que cada aresta conecta dois nós, chamados pontos finais da aresta (KREYSZIG, 1997). Assim, podemos escrever:

$$G = (V, E) \tag{16}$$

Diz-se que um vértice é incidente a uma aresta quando os dois estão conectados. Diz se que dois vértices são adjacentes quando ambos estão conectados por uma mesma aresta. O grau de um nó e dado pelo número de arestas incidentes a ele (KREYSZIG, 1997). Um nó ou uma aresta podem ser descritos através de números. Uma aresta pode ser descrita pelos nós a ela ligados: (i, j) onde i e j são os respectivos vértices conectados à aresta.

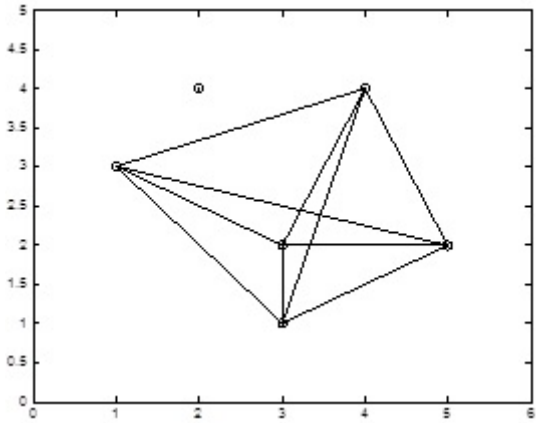

Figura 18: Exemplo de grafo com nó isolado em (2;4). Fonte: Próprio autor.

Um grafo formado a partir de um subconjunto de elementos de outro grafo é dito um **subgrafo** deste. Um **caminho** é um subgrafo tal que: i) à exceção de dois nós, todos os outros estão conectados a dois ramos; ii) em cada um desses dois nós, chamados de

nós terminais está conectado apenas um ramo; iii) nenhum subgrafo deste subgrafo contendo os mesmos nós terminais obedece às condições i) e ii) (BALABANIAN; BICKART, 1981).

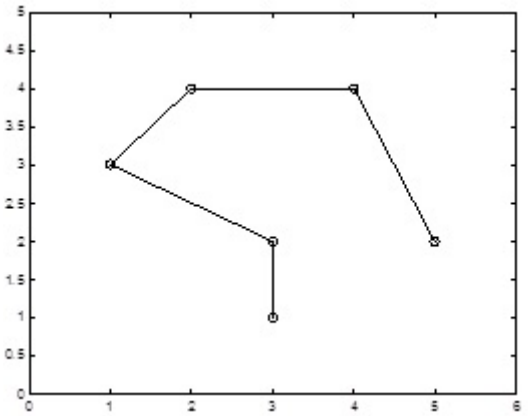

Figura 19: Exemplo de um caminho. Fonte: Próprio autor.

Um grafo é dito **conexo** se existe pelo menos um caminho entre quaisquer dois nós. Um **ciclo**, às vezes chamado de circuito, é um subgrafo cujos nós são conectados a dois ramos. Se os nós terminais de um caminho não são distintos e são coincidentes o resultado é um caminho fechado simples, que é um ciclo. Um ciclo pode ser descrito apenas utilizando seus ramos e, se ele não for um multigrafo (com ramos paralelos), ele também pode ser representado pelos seus nós em sequência (BALABANIAN; BICKART, 1981). Um **grafo completo** é aquele em que todos os nós são adjacentes (WEST, 2001).

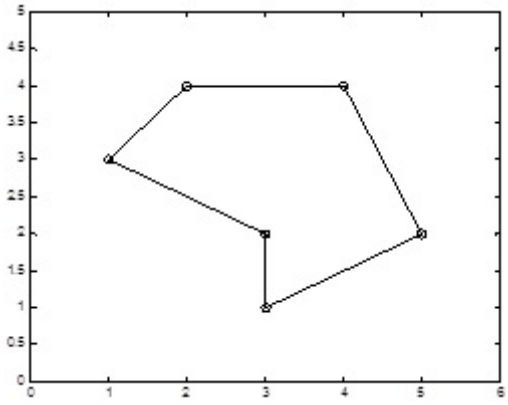

Figura 20: Exemplo de um grafo conexo cíclico. Fonte: Próprio autor.

Uma **árvore** é definida como um subgrafo conexo e acíclico. Ao descrever uma árvore, apenas listar os ramos é suficiente. Os ramos de uma árvore são chamados de *twigs*. Os nós conectados apenas a um *twig* são chamados de folhas da árvore. Os ramos de um grafo que não fazem parte de uma árvore são chamados de *links*. Complemento uma árvore é um subgrafo composto pelos *links* e os respectivos nós. Portanto, esses dois subgrafos complementares compartilham alguns nós e a união destes é o grafo. Em geral, a decomposição de um grafo em árvore e seu complemento não é única. Seja n o número de *twigs*. O número de nós de uma árvore será sempre $n+1$. Seja b o número de ramos do grafo, então o número de *links* é $b-n$ (BALABANIAN; BICKART, 1981). Em algumas aplicações, as árvores possuem estrutura hierárquica, com um nó no topo, chamado de

raiz (WILSON, 1996). Uma **árvore geradora** ou árvore completa é um subgrafo conexo de um grafo conexo contendo todos os nós do grafo, mas sem ciclos.

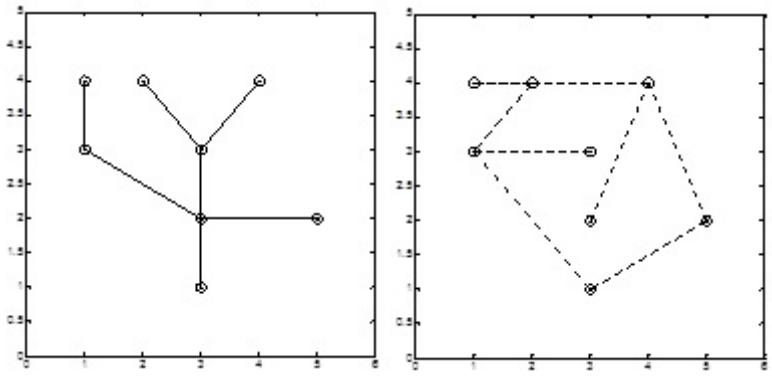

Figura 21: Exemplo de árvore (à esquerda) e um complemento desta árvore (à direita). Fonte: Próprio autor.

Se o grafo não é conexo, o objeto que correspondente à árvore é chamado de **floresta**, que é definida como a união das árvores de cada parte do grafo. A definição de **parte** de um grafo é um subgrafo conexo que, ao incorporar um ramo do grafo gera um subgrafo não conexo. O número de nós de um grafo contendo p partes é descrito como $n + p$. O complemento de uma floresta é normalmente chamado de coforest (BALABANIAN; BICKART, 1981).

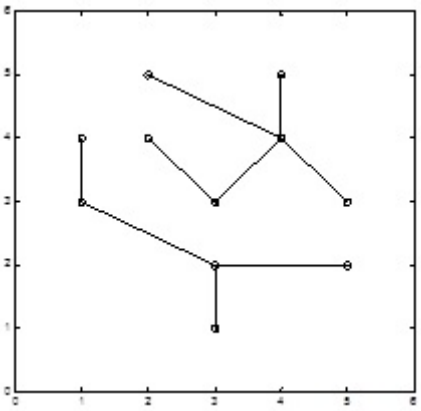

Figura 22: Exemplo de floresta. Fonte: Próprio autor.

A.2 Representação de Grafos

Uma maneira de representar um grafo é por meio de uma **matriz de adjacência**. Essa matriz representa as relações de adjacência entre nós e podem ser usadas para representar grafos e dígrafos (KREYSZIG, 1997). Neste texto, escolheremos a letra D para representar a matriz de adjacência. Para um grafo, os elementos $[d_{ij}]$ de uma matriz de adjacência são definidos por:

- $d_{ij} = 1$ se há uma aresta conectando os nós i e j;
- $d_{ij} = 0$ se não há conexão entre os nós i e j.

A partir dessa definição, percebe-se que a matriz D será uma matriz simétrica. Exemplo (KREYSZIG, 1997):

$$D = \begin{array}{c} \textit{nó} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix} \end{array} \tag{17}$$

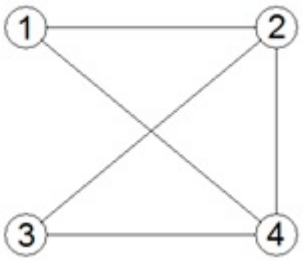

Figura 23: Exemplo de um grafo definido pela matriz de adjacência em eq:adja. Adaptado de (KREYSZIG, 1997).

Já para um dígrafo, os elementos da matriz de adjacência D são definidos da seguinte forma: $d_{ij} = 1$ se há uma aresta na direção de i para j; $d_{ij} = 0$ se não há conexão entre os nós i e j. O que resulta em uma matriz *assimétrica*. Exemplo (KREYSZIG, 1997):

$$D = \begin{array}{c} \frac{\textit{Para o nó}}{\textit{Do nó}} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \end{array} \tag{18}$$

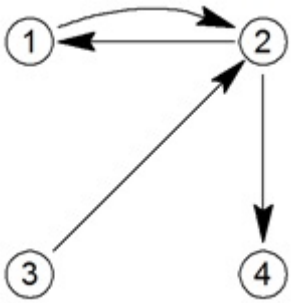

Figura 24: Exemplo de um grafo definido pela matriz de adjacência em (18). Adaptado de (KREYSZIG, 1997). Adaptado de (KREYSZIG, 1997).

Os grafos podem ainda ser representados de uma maneira mais compacta através das **listas de incidência de vértice** e **listas de incidência de aresta**. A primeira relaciona o as arestas incidentes a cada nó, enquanto a segunda mostra quais são os *endpoints* de cada aresta (KREYSZIG, 1997). Aqui, tanto os nós quanto as arestas serão representados por números, diferentemente da notação adotada em (KREYSZIG, 1997).

Em geral, listas são adotadas como forma de representação de grafos esparsos, ou seja, quando há poucas arestas. Além disso, as listas facilitam a execução de tarefas tais como ordenamento, classificação de maneira mais eficiente (KREYSZIG, 1997).

Tabela 15: Exemplo de lista de incidência de vértice.

Nó	Arestas Incidentes
1	1;5
2	1;2;3
3	2;4
4	3;4;5

Tabela 16: Exemplo de lista de incidência de aresta.

Aresta	Nós
1	1;2
2	2;3
3	2;4
4	3;4
5	1;4

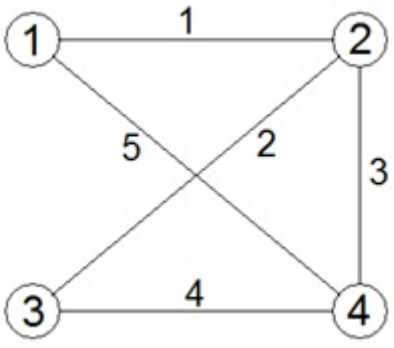

Figura 25: Grafo representado pelas listas de 15 ou 16. Adaptado de (KREYSZIG, 1997).

A **matriz de incidência** é uma representação de um grafo orientado , relacionando os nós aos ramos orientados do grafo. Se ela for a representação de todos os elementos do grafo, é denotada por A_c e chamada matriz de incidência completa. Sendo b o número de ramos e $n + p$ o número de nós, ela possui dimensão $(n + p) \cdot b$ (BALABANIAN; BICKART, 1981). Esta representação, em geral, produz uma matriz maior do que a matriz de adjacência (KREYSZIG, 1997). Dependendo da orientação e incidência entre os nós, os elementos desta matriz podem assumir os seguintes valores: $a_{ij} =$ +1 se o ramo j sai do nó i; $a_{ij} =$ -1 se o ramo j entra no nó i; $a_{ij} = 0$ se o ramo j não incide sobre o nó i.

Exemplo:

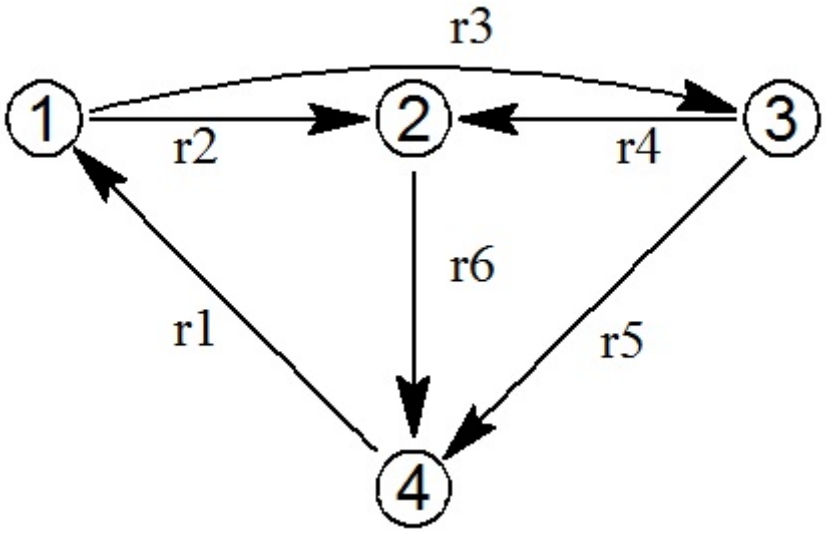

Figura 26: Um grafo orientado conexo com apenas uma partição. Adaptado de (KREYSZIG, 1997).

$$D = \begin{array}{c} \frac{Aresta}{Nó} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \begin{pmatrix} -1 & 1 & 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & -1 & 0 & 1 \\ 0 & 0 & -1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & -1 & -1 \end{pmatrix} \end{array} \tag{19}$$

A.3 Problema da Árvore Geradora Mínima

A árvore geradora mínima T em um grafo G (cujas arestas possuem comprimento $l_{ij} > 0$) é aquela cuja soma do comprimento das arestas l_{ij} é a menor dentre todas as árvores geradoras de G. Um algoritmo simples para resolução desse problema é o **algoritmo Guloso de Kruskal**. Esse algoritmo é particularmente adequado para grafos esparsos (KREYSZIG, 1997).

Um algoritmo é considerado guloso quando o mesmo é dito pouco sofisticado e capaz de gerar soluções ótimas apenas localmente. No entanto, este é capaz de gerar ótimos globais (WEST, 2001).

Outro algoritmo bastante utilizado para geração das árvores geradoras mais curtas é o **algoritmo de Prim**. Este algoritmo fornece uma árvore T em cada etapa, uma propriedade que o algoritmo de Kruskal não possui. No algoritmo de Prim, começando a partir

Algoritmo 5: Algoritmo guloso de Kruskal para obtenção de árvore geradora mínima.

Dados: $G = (V, E)$, l_{ij} para todas arestas de E.
Entrada: Arestas (i, j) de G e comprimentos l_{ij}.
Saída: Árvore geradora mais curta T em G.

1. Ordenar arestas de G em ordem crescente de comprimento.
2. Escolhe-las nesta ordem como arestas de T, rejeitando uma aresta apenas se ela forma um ciclo com arestas previamente escolhidas. Se $n-1$ arestas tiverem sido escolhidas, então fornecer T e parar.

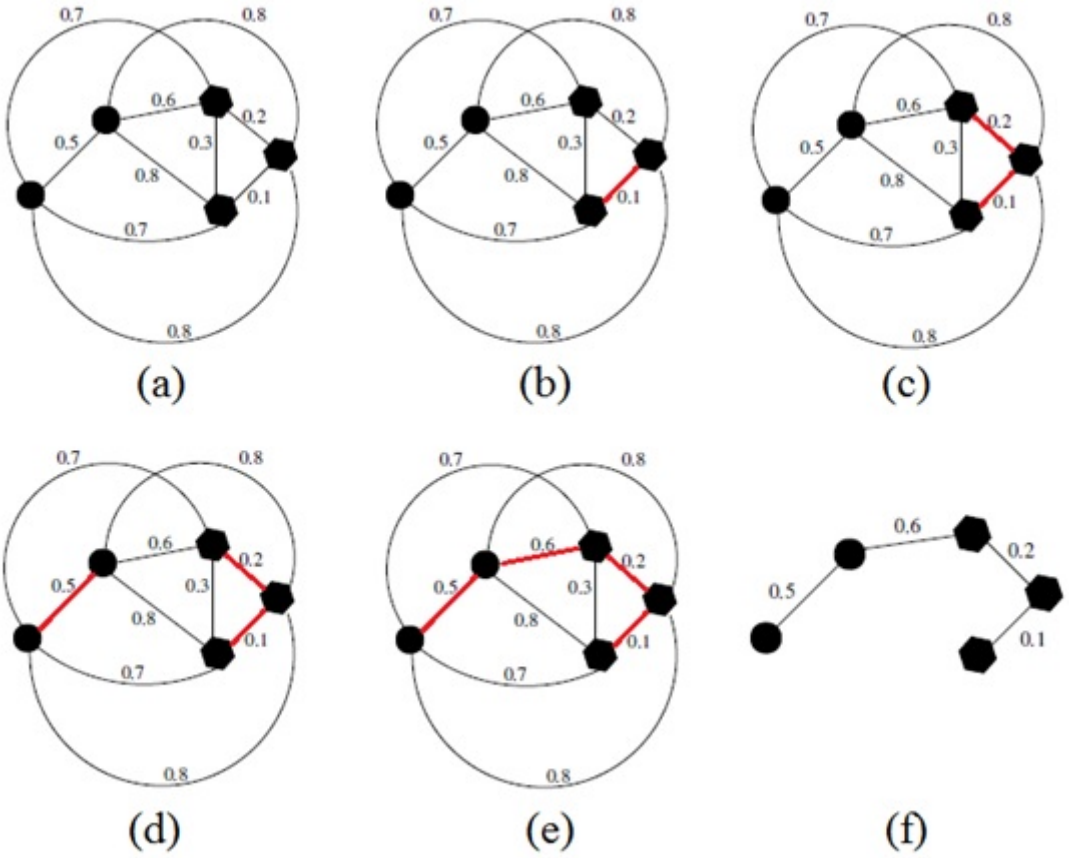

Figura 27: Exemplo de utilização do algoritmo de Kruskal. Em (a) temos o grafo completo, e em (f) chega-se à árvore geradora mínima. A aresta de peso 0.6 une dois nós de classes diferentes, logo estes nós são os protótipos ótimos. Adaptado de (PAPA et al., 2007).

de qualquer vértice, chamado 1, a árvore "cresce"adicionando-se arestas a ela, uma por vez, de acordo com alguma regra, até que T finalmente se torne uma árvore geradora com a característica de ser a mais curta. Chamamos de U o conjunto de vértices da árvore crescente T e de S o conjunto das suas arestas. Logo, inicialmente, $U = \{1\}$ e $S = \{\emptyset\}$ e no final $U = V$.

Algoritmo 6: Algoritmo de Prim para obtenção de árvore geradora mínima.

Dados: Grafo conexo G com vértices 1, 2, ..., $n+1$ e arestas (i, j) de comprimento $l_{ij} > 0$.
Entrada: $n + 1$,arestas (i, j) de G e comprimentos l_{ij}.
Saída: Conjunto de arestas S de uma árvore geradora mais curta T em G. Comprimento $L(T)$.

1. **Passo inicial**: Fazer $i(k) = 1$, $U = \{1\}$ e $S = \{\emptyset\}$. Rotular vértice $k = (2, ..., n + 1)$ com $\lambda_k = l_{ik}$ ($= \infty$ se $(1, k)$ não existir em G).
2. **Adição de um ramo à árvore** T. Seja λ_j o menor λ_k para k não pertencente a U. Incluir vértice j em U e aresta $(i(j), j)$ em S. Se $U = V$, calcular $L(T) = \Sigma l_{ij}$. Parar. Caso contrário, continuar.
3. **Atualização de rótulos**: Para cada k não pertencente a U, se $l_{jk} < \lambda_k$, então fazer $\lambda_k = l_{jk}$ e $i(k) = j$. Ir para o passo 2.

Logo, no primeiro passo, os rótulos λ_2, λ_3, ..., λ_{n+1} dos vértices 2, ..., $n + 1$ são os comprimentos das arestas que os conectam a 1. No passo 2, se escolhe o mais curto deles como a primeira aresta de T e se inclui o seu outro *endpoint* a U (escolhendo o menor j se houver muitos, para tornar o processo único). A atualização de rótulos de 3 concerne aqueles nós ainda não pertencentes a U. Vértice k inicia esta etapa com $\lambda_k = l_{i(k),k}$. Se $l_{jk} < \lambda_k$, isso quer dizer que k está mais próximo do novo membro de U, j, do que do seu antigo "vizinho mais próximo"$i(k)$ em U. Então se atualiza o rótulo de k substituindo $\lambda_k = l_{i(k),k}$ por $\lambda_k = l_{j,k}$ e fazemos $i(k) = j$. Se, no entanto, $l_{j,k} \geq \lambda_k$ (o antigo rótulo de k) então não se muda o rótulo de k. Logo, o rótulo de k sempre identifica o seu vizinho mais próximo em U, e isso é atualizado no passo 3 a cada vez que U e a árvore crescem. A partir dos rótulos finais, pode-se regredir até se ter toda a árvore e a partir dos seus valores numéricos se calcula o comprimento total desta árvore.