

UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE CIÊNCIAS EXATAS E NATURAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Reginaldo Cordeiro dos Santos Filho

**Identificação de *Binder* Através de Medições
Fantasmas Utilizando Técnicas de Aprendizado
de Máquina**

Belém-PA

2016

Reginaldo Cordeiro dos Santos Filho

**Identificação de *Binder* Através de Medições
Fantasmas Utilizando Técnicas de Aprendizado
de Máquina**

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal do Pará, como requisito parcial para a obtenção do Grau de Mestre em Ciência da Computação na área de concentração: Sistemas de Computação. Linha de Pesquisa: Sistemas Inteligentes.

Belém-PA

2016

Dados Internacionais de Catalogação-na-Publicação (CIP)
Sistema de Bibliotecas da UFPA

Santos Filho, Reginaldo Cordeiro dos, 1988-
Identificação de binder através de medições
fantasmas utilizando técnicas de aprendizado de máquina
/ Reginaldo Cordeiro dos Santos Filho. - 2016.

Orientador: Claudomiro de Souza de Sales
Júnior.

Dissertação (Mestrado) - Universidade
Federal do Pará, Instituto de Ciências Exatas e
Naturais, Programa de Pós-Graduação em Ciência
da Computação, Belém, 2016.

1. Telecomunicações. 2. Sistemas
telefônicos. 3. Inovações tecnológicas. 4.
Identificação de binders. 5. Medição em circuito
fantasma. I. Título.

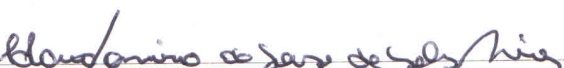
CDD 22. ed. 621.38216

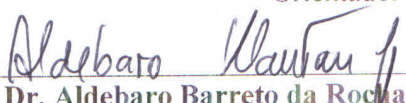
UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE CIÊNCIAS EXATAS E NATURAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

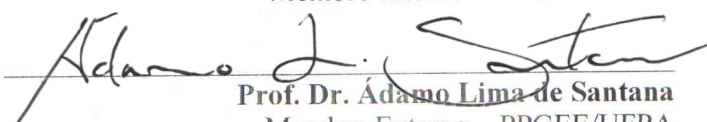
REGINALDO CORDEIRO DOS SANTOS FILHO

**IDENTIFICAÇÃO DE *BINDER* ATRAVÉS DE MEDIÇÕES FANTASMAS
UTILIZANDO TÉCNICAS DE APRENDIZADO DE MÁQUINA**

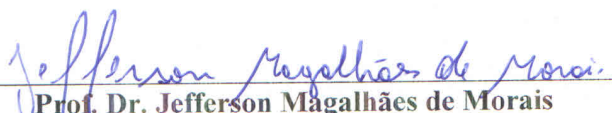
Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal do Pará como requisito para obtenção do título de Mestre em Ciência da Computação, defendida e aprovada em 22/02/2016, pela banca examinadora constituída pelos seguintes membros:


Prof. Dr. Claudomiro de Souza de Sales Júnior
Orientador – PPGCC/UFPA


Prof. Dr. Aldebaro Barreto da Rocha Klautau Júnior
Membro Interno – PPGCC/UFPA


Prof. Dr. Adamo Lima de Santana
Membro Externo – PPGEE/UFPA

Visto:


Prof. Dr. Jefferson Magalhães de Moraes
Coordenador do PPGCC/UFPA

Prof. Dr. Jefferson Magalhães de Moraes
Coordenador do PPGCC
Mat.: SIAPE: 2378314

*Dedico esta Dissertação aos meus três pilares: avó
Tirzach Lourenço, mãe Ilza Anete Lourenço
dos Santos e tia Ilza Lourenço;
amigos e conhecidos.*

Agradecimentos

À minha família, Avó Tirzach Lourenço, Mãe Ilza Anete Lourenço dos Santos e Tia Ilza Lourenço por acreditar e investir em mim.

À minha noiva pelo companheirismo, cumplicidade e carinho.

Ao meu orientador Prof. Dr. Claudomiro Sales por toda orientação no período de mestrado e por ter confiado a mim o assunto tratado aqui neste documento.

A todos os companheiros do Laboratório de Eletromagnetismo Aplicado (Lea) que ao longo do percurso do mestrado me proporcionaram triunfantes momentos de sabedoria.

Aos meus amigos e conhecidos por compartilharem momentos de felicidade, tristezas, dores, conhecimentos, diversões, descontrações, e principalmente todas as críticas construtivas de cunho nobre.

“O homem está condenado a ser livre.”

Jean-Paul Charles Aymard Sartre

*“A natureza não é cruel, apenas implacavelmente
indiferente. Essa é uma das lições mais duras
que os humanos têm de aprender.”*

Clinton Richard Dawkins

Resumo

Com o advento da internet os sistemas de telecomunicações evoluíram no sentido de prover altas taxas de transmissão de dados em condutores de cobre. Atualmente, vivenciamos a 5^a geração de banda larga desenvolvida em cima da rede antiga de telefonia e podendo alcançar até 10 Gb/s de taxa agregada em pares trançados de cobre. A era Gb em cobre é uma estratégia interessante no sentido de reaproveitar uma estrutura de telefonia já existente e mundialmente implantada, ao passo de que substituir totalmente os condutores de cobre por fibras óptica ainda é uma operação onerosa. Muito se investe em tecnologias para aumentar com qualidade a taxa de transmissão em pares trançados. Linhas coordenadas por *bonding* e livre de ruídos por *vectoring* garantem bons rendimentos do uso do cobre em sistemas híbridos fibra-cobre implantados atualmente. Seguindo esta perspectiva, este trabalho apresenta uma proposta de estender os princípios da chamada Qualificação de Enlaces telefônicos, da qualificação de um enlace individual para a qualificação de *binders* e cabos. A proposta inclui um método de identificação de *binders* de pares trançados de cobre. O método se baseia na análise de medições de circuitos fantasmas. Este tipo de circuito é frequentemente usado para melhorar as taxas de transmissão de dados em sistemas de comunicação, mas neste trabalho, esses circuitos serão utilizados para revelar se 4 condutores compostos por 2 pares trançados estão próximos o suficiente para serem considerados no mesmo *binder*. O método ainda consegue estimar o comprimento em que dois pares trançados compartilham o mesmo *binder*, possibilitando descobrir o local onde os pares trançados se dividem em cabos diferentes. A identificação de *binders* é feita através de técnicas de reconhecimento de padrão, máquina de vetores de suporte, *K*-means e modelo de misturas de Gaussianas, aplicadas às características extraídas do parâmetro S_{11} obtido pela medição do sinal no domínio do tempo e da frequência de dois pares trançados que formam o circuito fantasma. Através de medições fantasmas reais realizadas em laboratório, obteve-se diferentes resultados possibilitando comparar a performance de cada técnica de aprendizado de máquina empregada ao problema. Os resultados alcançaram níveis aceitáveis de acurácia para o método proposto.

Palavras-chave: Identificação de *Binders*, DSL, Medição em Modo Fantasma, Máquina de Vetores de Suporte, *K*-means, GMM.

Abstract

The advent of the world wide web has made the telecommunications systems evolve in order to provide high data rates on copper conductors. Today, we experience a 5th generation broadband copper access developed on the plain old telephone network and may reach up to 10 Gb/s aggregate rate in twisted copper pairs. The Gb era over copper is an interesting strategy to reuse a plain old telephone structures that were globally deployed, while replacing them with optical fibers is still a costly decision. There are investments in technologies to increase the quality of the transmission rate in twisted pairs. Lines coordinated by bonding and noise-free by vectoring ensure good yields of copper use in hybrid fiber-copper systems. Following this perspective, this work presents a proposal to extend the principles of the so called Loop Qualification, from the qualification of an individual loop to qualifying binders and cables. The proposal includes a binder identification method of twisted copper pairs. The method is based on analysis of phantom circuits measurements. This type of circuit is often used to improve data transmission rates in telecommunications systems, but in this work, phantoming is used to reveal if a 4-wire loop composed by two twisted pairs are close enough and well balanced in order to be considered in the same binder. The method can also estimate the length in which two twisted pairs share the same binder, enabling discover where the twisted pairs split into different cables. The identification is done via application of pattern recognition techniques, support vector machines, K -means and Gaussian mixture model, applied to the features extracted from the S_{11} parameter of the signal measurement in the frequency and time domain of two twisted pairs forming the phantom circuit. Through phantom measurements performed in the laboratory was obtained different results enabling to compare the performance of each machine learning algorithm used in the problem. The results achieved acceptable levels of accuracy for the proposed method.

Key-words: Binder Identification, DSL, Phantom-Mode Measurement, Support Vector Machine, K -means, GMM.

Lista de ilustrações

Figura 1 – Representações de cabo: (a) ilustração de um cabo e seus <i>binders</i> internos; (b) cabo real e seus <i>binders</i> internos.	16
Figura 2 – Distância de compartilhamento em que pares trançados compartilham o mesmo cabo.	17
Figura 3 – Mapa espectral de serviços xDSL em linhas de telefone e coexistência com a tecnologia <i>FemtoWoC</i> . Extraído de: (1).	20
Figura 4 – <i>FemtoWoC</i> sujeito a interferência de outras linhas no mesmo <i>binder</i> trafegando serviços xDSL. Adaptado e extraído de: (1).	20
Figura 5 – Evolução de tecnologias de banda larga em fios de cobre e a relação de distância entre o ponto de distribuição e a casa do cliente em sistemas híbrido fibra-cobre. Extraído de: (2).	21
Figura 6 – Visão transversal de dois cabos contendo três <i>binders</i> cada um. Cada <i>binder</i> contém sete pares trançados.	27
Figura 7 – Circuito com sinalização em modo comum.	28
Figura 8 – Circuito com sinalização em modo diferencial.	29
Figura 9 – Interferência em modo comum em um sistema com sinalização em modo diferencial.	29
Figura 10 – Circuito fantasma utilizando dois pares trançados. Os <i>baluns</i> b são transformadores utilizados para fazer com que parte do sinal fantasma trafegue no modo comum de cada par trançado.	30
Figura 11 – Circuito com transmissão em modo fantasma.	31
Figura 12 – Candidatos a hiperplanos de separação e hiperplano ótimo em destaque.	35
Figura 13 – Exemplo de uma função kernel $k(\cdot, \cdot)$ para separação de classes em problemas não lineares.	36
Figura 14 – Exemplo de um hiperplano com margem maximizadas em uma base de dados com duas classes. As amostras circuladas são vetores de suporte.	37
Figura 15 – Exemplo do resultado de uma clusterização utilizando o algoritmo <i>K</i> -means.	39
Figura 16 – Exemplo de uma validação cruzada com $k = 4$ aplicada a uma base de dados.	44
Figura 17 – Gráfico contendo 45 medições em modo fantasma. As curvas de com linhas no estilo ponto tracejadas representam as medições de PTs de diferentes cabos, as curvas com linhas no estilo tracejadas são medições de <i>binders</i> diferentes e as curvas de linhas cheias são medições de mesmo <i>binder</i>	46

Figura 18 – Visão transversal de um cabos contendo três <i>binders</i> cada um. No exemplo, um circuito fantasma (CF 1) é formado entre dois PTs que estão em <i>binders</i> diferentes (A e B). Outro circuito fantasma (CF 2) é formado por dois PTs que estão no mesmo <i>binder</i> (C).	47
Figura 19 – Semelhança entre duas medições em modo fantasma oriundas de cenários diferentes. Em azul, dois PTs no mesmo <i>binder</i> e, em vermelho, dois PTs em <i>binders</i> diferentes.	47
Figura 20 – Variância do período extraída no domínio da frequência de um sinal fantasma formado por dois PTs de 500 m de comprimento que estão situados no mesmo <i>binder</i>	49
Figura 21 – Variância da magnitude extraída no domínio da frequência de um sinal fantasma formado por dois PTs de 500 m de comprimento que estão situados no mesmo <i>binder</i>	50
Figura 22 – Linhas espectrais aplicando PSD no domínio da frequência de uma medição em modo fantasma.	51
Figura 23 – Primeiro ponto da TDR de um sinal fantasma formado por dois PTs de 500 m de comprimento que estão situados no mesmo <i>binder</i> . A quarta característica tem relação direta com o primeiro ponto S_{11}^{PM} , que tem relação direta com efeito NER. O círculo representa o nível NER.	52
Figura 24 – Fluxograma do algoritmo da identificação de <i>binders</i>	53
Figura 25 – Vetor característico de 90 amostras S_{11}^{PM} da base de dados considerando um problema de identificação binário de <i>binders</i> . Mesmo <i>binder</i> e <i>binders</i> diferentes são consideradas como apenas uma classe.	54
Figura 26 – Visualização da f_4 para 90 amostras S_{11}^{PM} selecionadas randomicamente da base de dados. Uma separação clara das duas classes pode ser vista nesta figura.	55
Figura 27 – As mesmas 90 amostras S_{11}^{PM} . Agora considerando a identificação de <i>binders</i> como um problema ternário com 30 amostras de cada classe.	55
Figura 28 – Vetor característico de 30 medições S_{11}^{PM} para cada classe: (parte superior) Destaca a característica f_2 apenas. (parte inferior) Destaca a característica f_4 apenas.	56
Figura 29 – Representação visual do algoritmo de mapeamento utilizado para rotular <i>clusters</i>	58
Figura 30 – Dois PTs compartilhando o mesmo cabo e posteriormente se dividindo em cabos diferentes. O ponto em que dois PTs se dividem tem comportamento de circuito aberto. A distância de compartilhamento representa o comprimento que eles compartilham o mesmo cabo.	62
Figura 31 – Exemplo de singularidades encontradas em um sinal em modo fantasma.	64
Figura 32 – Exemplo da aplicação do filtro de bordas em um sinal em modo fantasma.	64

Figura 33 – A resposta no domínio do tempo para dois PTs no mesmo <i>binder</i> . O círculo e o quadrado são os pontos de subida e descida do pulso, respectivamente.	66
Figura 34 – Duas medições S_{11}^{PM} : PTs no mesmo <i>binder</i> (linha cheia) têm alta periodicidade $ S_{11}^{PM} $ e PTs em <i>binders</i> diferentes (linha tracejada) têm baixa periodicidade.	66
Figura 35 – <i>Setup</i> para medição fantasma de uma porta. O <i>Network Analyzer</i> é controlado pelo computador e gera sinal em modo comum que é convertido para sinal em modo diferencial conectado em paralelo em ambos os PTs.	69
Figura 36 – Hiperplano para identificação de <i>binders</i> e vetores de suporte na base de treino.	75
Figura 37 – Resultado da clusterização do algoritmo <i>K</i> -means para o melhor classificador de combinação [1,4] ($\sigma_p^2, \mathcal{A}_1$) aplicado à base de treinamento. Os <i>clusters</i> foram devidamente rotulados através da técnica de rotulação descrita neste trabalho.	78
Figura 38 – Resultado da clusterização do algoritmo <i>K</i> -means para o melhor classificador de combinação [4] (\mathcal{A}_1) aplicado à base de treinamento. Os <i>clusters</i> foram devidamente rotulados através da técnica de rotulação descrita neste trabalho.	78
Figura 39 – Resultados do GMM para o melhor classificador usando apenas a característica \mathcal{A}_1 na base de treinamento.	80
Figura 40 – Probabilidade posterior de cada amostra da base de treinamento. Poucas interseções de curvas indicam um classificador coerente para classificação de dados.	81
Figura 41 – Gráfico em barras para os três cenários distintos. Margem de erro em diferentes visões: mesmo <i>binder</i> (2,44%), <i>binders</i> diferentes (1,2%) e todas as medições (1,35%). O intervalo de confiança usado foi de 95%.	85

Lista de tabelas

Tabela 1	– Rotulação imediata de acordo com a intensidade do efeito NER.	58
Tabela 2	– Mapeamento e rotulação através do método proposto.	59
Tabela 3	– Lista de variáveis.	60
Tabela 4	– Média de acurácia para cada combinação de características e classe do algoritmo SVM. <i>CPIC</i> e <i>CPIB</i> significam, respectivamente, Características para o PIC e <i>Binder</i> no processo. <i>MB</i> , <i>BD</i> e <i>CD</i> significam a taxa de acerto por classe: Mesmo <i>Binder</i> , <i>Binders</i> Diferentes e Cabos Diferentes.	71
Tabela 5	– Frequência absoluta considerando o aparecimento de característica um a um apenas.	72
Tabela 6	– Frequência absoluta considerando o aparecimento de característica dois a dois.	72
Tabela 7	– Frequência absoluta considerando o aparecimento de característica três a três.	73
Tabela 8	– Acurácia do algoritmo SVM para cada classe usando a característica $[\mathcal{A}_1]$ para o PIC e $[\sigma_p^2, n_\phi, \mathcal{A}_1]$ para o PIB.	74
Tabela 9	– Média de acurácia para cada combinação e classe pelo algoritmo <i>K</i> -means.	76
Tabela 10	– Média de acurácia das melhores combinações de cada conjunto $C(n, i)$ para algoritmo <i>K</i> -means.	76
Tabela 11	– Média de acurácia para cada classe usando as características $[\sigma_p^2, \mathcal{A}_1]$ para a identificação de cabo e <i>binder</i>	77
Tabela 12	– Média de acurácia para cada classe usando a característica $[\mathcal{A}_1]$ para a identificação de cabo e <i>binder</i>	77
Tabela 13	– Média de acurácia para cada combinação e classe pelo algoritmo GMM.	79
Tabela 14	– Média de acurácia das melhores combinações de cada conjunto $C(n, i)$ para algoritmo GMM.	79
Tabela 15	– Acurácia do GMM para cada classe usando a característica \mathcal{A}_1 na identificação de <i>binder</i>	79
Tabela 16	– Visão geral da base de dados para estimação de comprimento.	82
Tabela 17	– Comprimentos estimados e erro de todas as 164 medições S_{11}^{PM}	83
Tabela 18	– Distribuição de erro (%)	84
Tabela 19	– Erro por intervalo (%)	84
Tabela 20	– Erro médio proporcional separado por comprimento.	84

Sumário

1	Introdução	15
1.1	Trabalhos Relacionados	18
1.2	Motivação	19
1.3	Justificativa	22
1.4	Objetivo Geral	23
1.5	Objetivos Específicos	23
1.6	Estrutura da Dissertação	24
2	Transmissão em Modo Fantasma	26
2.1	Terminologias e Contextualização do Trabalho	26
2.1.1	Par Trançado de Cobre	26
2.1.2	<i>Binder</i>	26
2.1.3	Cabo	26
2.2	Modo de Sinalização	28
2.2.1	Modo Comum	28
2.2.2	Modo Diferencial	28
2.3	Transmissão em Modo Fantasma	29
2.4	Qualificação de Enlace	31
2.5	Identificação de Topologia do Enlace	32
3	Aprendizado de Máquina na Identificação de <i>Binders</i>	34
3.1	Máquina de Vetores de Suporte	34
3.2	<i>K</i> -means	38
3.3	Modelo de Misturas de Gaussianas	39
3.4	SVM, <i>K</i> -means e GMM na Identificação de <i>Binders</i>	41
3.5	Validação Cruzada Estratificada	43
4	Método de Identificação de <i>Binder</i>	45
4.1	Considerações Sobre Características do Sinal Fantasma	45
4.2	Extração de Características do Sinal Fantasma	48
4.3	Algoritmo para a Identificação de <i>Binders</i>	52
4.4	Análise das Características	54
4.5	Método de Rotulação de <i>Clusters</i>	56
4.6	Identificação do Comprimento de Compartilhamento	62
5	Resultados	68
5.1	Cenário de Medição, Coleta e Transformação dos Dados	68
5.2	Resultados da Máquina de Vetores de Suporte	70
5.3	Resultados do <i>K</i> -means	74
5.4	Resultados do Modelo de Misturas de Gaussianas	77

5.5 Resultados da Estimação de Comprimento	81
Conclusão	86
Publicações	88
Referências	89

1 Introdução

Tecnologias modernas, tais como *G.fast* (acrônimo para *Fast Access to Subscriber Terminals*), *XG.fast* e *Femto WoC*, estão se tornando realidade gradativamente. A primeira, especificada pela ITU-T (3, 4) e amplamente discutida na literatura (5, 6, 7, 8), considerada a quarta geração de banda larga, trabalha em taxa agregada de até 1 Gb/s (*downstream* e *upstream*) em Pares Trançados (PTs) de cobre a uma distância de até 250 m da rede de distribuição, e com estimativa de trabalhar com taxas de até 2 Gb/s em 2016 (2). A segunda, considerada como o início da quinta geração de banda larga, promete taxas de 10 Gb/s (*downstream* e *upstream*) em PTs de cobre a uma distância de até 70 m da rede de distribuição (2, 9). A terceira, com trabalhos em voga (1, 10), torna possível a utilização de tecnologia de cobre para dar suporte a *femtocells* e prover serviços de transmissão de dados banda larga.

Todas essas tecnologias têm potencial de mercado para suprir as demandas crescentes de transmissão de dados de novos serviços multimídia que surgem a cada dia. No entanto, para usufruir dessas tecnologias com altas taxas de transmissão no cobre, deve-se aumentar a largura de banda, e.g., até 212 MHz para *G.fast* e até 500 MHz para *XG.fast*. Dificilmente uma nova tecnologia nasce pronta, o que implica dizer que um conjunto de problemas precisam ser tratados paralelamente ao processo de desenvolvimento das próprias tecnologias. Tratamento de ruído conhecido como *crosstalk cancellation* ou *Vectoring* (11, 12), investigação e análise da instabilidade do canal em ambientes *G.fast* (7), assim como o estudo da eficiência energética em ambientes *G.fast* (13), são alguns dos estudos que se debruçam em resolver ou mitigar problemas oriundos dessas tecnologias emergentes.

O gerenciamento das linhas de transmissão mais adequadas a receber essas novas tecnologias, redução de ruídos de fontes externas, gerenciamento de espectro, manutenção preditiva do sistema e determinar o serviço a ser implantado com o conhecimento da taxa de transferência máxima que pode ser oferecida, baseado na distância máxima do ponto de distribuição, também são algumas das tarefas que surgem com essas novas tecnologias e as formas eficientes de implantação de soluções também precisam ser discutidas.

Define-se cabo aqui como sendo uma estrutura formada por pares trançados que são agrupados em *binders*, como pode ser visto na Figura 1. O *binder* é um aglomerado de pares trançados de uma rede de telefonia. Em um conjunto de cabos, serviços diferentes que podem ser da mesma operadora ou de operadoras diferentes podem ter sobreposição de frequência afetando mutuamente a qualidade do serviço. Se esses serviços estão no mesmo cabo, ou no pior caso compartilham o mesmo *binder*, a interferência pode afetar severamente os serviços, podendo torná-los inoperante.

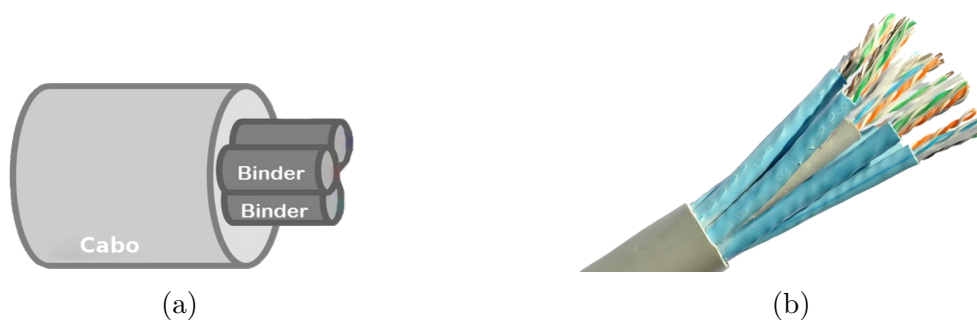


Figura 1 – Representações de cabo: (a) ilustração de um cabo e seus *binders* internos; (b) cabo real e seus *binders* internos.

A identificação de *binder* é um tópico relevante em pesquisas sobre sistemas de transmissão de dados que trabalham com pares trançados. Atualmente, a maioria das linhas de telefonia são heranças dos chamados *Plain Old Telephone Services* (POTSs). Conhecer a estrutura de uma rede de telefonia é essencial para que os operadores possam explorar ao máximo os novos serviços de transmissão de dados baseado em cobre, que podem ser oferecidos aos clientes. Entretanto, na maioria dos casos, as operadoras têm pouca informação da estrutura antiga das suas redes de cobre (14) e são essas redes que serão reusadas para implantar esses novos serviços.

Como essas tecnologias modernas usam extensas bandas espectrais, o gerenciamento dinâmico do espectro proporciona uma coordenação inteligente do sinal que viaja pelas diferentes linhas de transmissão. Mas para que seja possível aplicar o gerenciamento, é necessário que os operadores saibam quais pares trançados estão situados no mesmo *binder* (15). Além disso, o conhecimento sobre a topologia da rede e a distribuição dos pares em diferentes *binders* e cabos ajuda as operadoras a gerenciar a sua rede de transmissão e determinar o tipo de serviço que pode ser oferecido a um cliente específico. Para obter informações sobre a estrutura da rede é necessário aplicar técnicas de qualificação de enlace (16, 17, 18).

Adicionalmente, ter o domínio completo sobre a estrutura da rede de transmissão de cobre considerando a distribuição de pares trançados dentro de cada cabo também é útil para produzir diagnósticos. Quando uma falta ocorre em um par trançado, é interessante testar se os outros pares trançados do mesmo *binder* apresentam falta também. Além disso, diagnóstico de *binder* é mais confiável que diagnóstico de par trançado individual, pois os pares nos novos sistemas de transmissão de dados são coordenados para redução e eventualmente a eliminação de ruído devido a *crosstalk*.

A transmissão de sinais em modo fantasma é uma técnica usada para criar um novo canal de comunicação virtual sem necessidade de criar um canal físico (19). Essa tecnologia pode ser usada para melhorar as taxas de transmissão de dados em sistemas transmissão baseados em cobre e, como o padrão admite altas taxas de transmissão, é considerada

uma solução para prolongar o uso de tecnologias de fios de cobre, principalmente quando combinado com outras tecnologias de processamento de sinais, e.g., *vectoring* (11), pois aumenta o número disponível de canais usados para transmitir.

As técnicas utilizadas na área de identificação de topologias se preocupam na identificação de linhas individuais e quase nenhuma informação é dada sobre as outras linhas de clientes vizinhos. O presente trabalho expande o princípio da identificação de linha para a identificação de *binder* e cabo. O conhecimento sobre como os pares trançados estão distribuídos ao longo do cabo contribui para o gerenciamento das linhas de transmissão, e ainda, na predição de futuras implantações de tecnologias banda larga, e.g., redes de acesso híbridas com fibra e cobre. Além da identificação se pares compartilham o mesmo *binder*, é relevante também identificar a distância que os pares seguem juntos no mesmo *binder* ou cabo. Por isso, este trabalho também desenvolve um método para estimar a distância de compartilhamento.

A Figura 2 explica visualmente o que o método que calcula a estimativa do comprimento de compartilhamento é capaz de fazer. Nesta figura, tem-se a disposição 3 cabos e 4 *binders*, além de uma região com construções civis. Quaisquer que sejam os dois pares trançados, sendo do mesmo *binder* ou não, que estejam sendo usados para a transmissão em modo fantasma é possível identificar a distância que estes pares estão compartilhando o mesmo cabo 1, e depois dividem-se em cabos diferentes ou vão direto para as dependências do cliente. Essa informação é usada para determinar a quantidade de ruído entre linhas não coordenadas ou de sistemas diferentes compartilhando o mesmo cabo, determinando assim as taxas máximas que podem ser alcançadas (1).

O presente trabalho faz uso da transmissão em modo fantasma para construir a base de dados que será utilizada na identificação de *binder*.

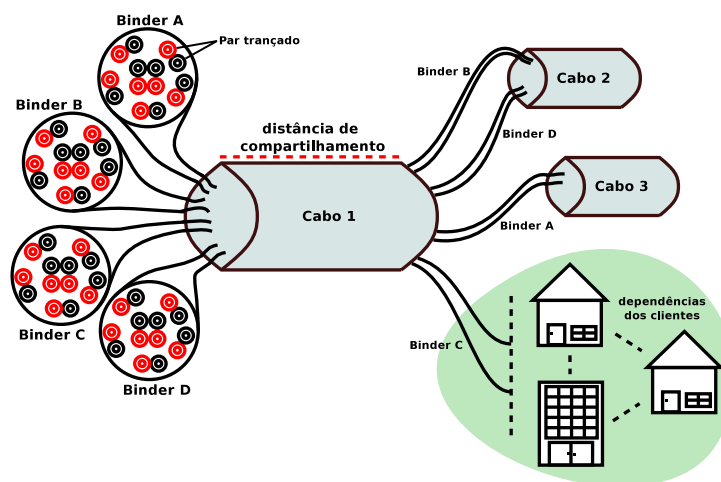


Figura 2 – Distância de compartilhamento em que pares trançados compartilham o mesmo cabo.

1.1 Trabalhos Relacionados

Alguns estudos apontam o sinal em modo fantasma como sendo a tecnologia que prolongará o uso de condutores de cobre (11, 20, 21, 22). Embora não seja uma tecnologia padronizada por uma entidade internacional, tem-se na literatura um avanço significativo de trabalhos que usam o modo fantasma como uma alternativa que auxilia no aumento de taxas de transmissão em condutores de cobre e destacam a hibridização de sistemas fibra-cobre nos próximos anos (20).

Em (23), é proposto um modelo físico para circuitos fantasma baseado na teoria das linhas de transmissão multicondutor, explorando a possibilidade de adicionar o circuito fantasma como um terceiro canal de comunicação. Já em (24), um estudo mais específico e um modelo similar é proposto para cabos *quad*. Esses modelos não são usados no trabalho em questão, pois consideram condições ideais quanto à homogeneidade do meio e distância constante entre os pares trançados. O trabalho (25) faz um estudo do modo de transmissão fantasma e uma avaliação sobre o desempenho que este modo pode oferecer. O *setup* construído para esta dissertação é baseado no *setup* desenvolvido no trabalho (25).

Os dois trabalhos encontrados na literatura que discutem a identificação de *binder* são destacados a seguir. O trabalho (26) apresenta um método para identificação de *binder* baseado em medidas de ruído da linha. Este método leva um tempo considerável para produzir resultados (alguns dias são necessários para identificar os pares trançados mais próximos). Já no trabalho (15) é feito um estudo que revela a possibilidade de identificação de *binder* através da interpretação de alguns parâmetros de medição como impedância de entrada e reflexão do sinal no domínio do tempo do circuito fantasma composto por dois pares trançados candidatos a identificação.

O estudo (15) mostra que circuitos fantasmas compostos por dois PTs em diferentes *binders* apresentam alta impedância característica, conseqüentemente criando um descasamento com a baixa impedância do equipamento de medição. Enquanto que circuitos fantasmas de PTs no mesmo *binder* têm baixa impedância característica e, portanto, boa parte do sinal se propaga ao longo da linha devido ao melhor casamento entre o circuito fantasma e o equipamento de medição. Essa dicotomia mostrada no trabalho (15) traz indícios da identificação de *binders*, i.e., anuncia uma possível identificação da origem dos pares trançados através da análise do comportamento do sinal fantasma em diferentes domínios. Apesar do trabalho discutir sobre a identificação de *binders*, não é apresentada uma proposta de aplicação para este fim. Ainda, o trabalho considera que existe blindagem entre os *binders*. Até onde é conhecimento do autor, não existem cabos com blindagem entre *binders*, mas apenas a blindagem que envolve todos os *binders*, i.e., a blindagem do cabo.

Além disso, a definição de *binder* descrita no trabalho não está clara. O compor-

tamento do sinal fantasma mostrado nos gráficos refere-se às medições que foram feitas em PTs no mesmo *binder* e medições em PTs de *binders* diferentes. Especificamente no segundo caso, as medições de *binders* diferentes assemelham-se a circuitos fantasmas formados por PTs situados em cabos diferentes. Partindo desta premissa, a identificação de *binders* (como é anunciada no trabalho) parece ser uma tarefa fácil de se realizar, uma vez que os padrões mostrados dos dois cenários são bem distintos. No entanto, acredita-se que, o que se obteve como conclusão do trabalho foi uma identificação de cabos (e não de *binders* como é anunciado).

Nesta dissertação, como contribuição adicional, um terceiro padrão pode ser identificado: PTs em *binders* diferentes, mas situados no mesmo cabo. Este padrão anuncia mais um nível de especificidade em relação ao mostrado no trabalho supracitado. A dissertação mostra ainda que é possível realizar a identificação automática de *binders* e cabos através de algoritmos de aprendizado de máquina. E ainda, pode-se mensurar o comprimento de cabo compartilhado por dois PTs.

1.2 Motivação

A agregação de taxas de transmissão de dados em pares trançados através da técnica *bonding* especificada pela ITU-T G.998.x (27, 28, 29), juntamente com o cancelamento de ruído através da técnica *vectoring* (11) e gerenciamento de *binders* (30), a utilização da transmissão em modo fantasma (19, 31), e ainda, o aperfeiçoamento de técnicas como *G.fast* (5) e o mais recente *XG.fast* (9) reforçam a utilização do par trançado de cobre na última milha em sistemas DSL (22). Paralelamente, há também estudos que se preocupam com a coexistência entre DSL e outras tecnologias como *Power Line Communication*, ou PLC (32, 33), o que reforça e ramifica ainda mais os estudos sobre tecnologias de cobre. Em época que se fala muito em utilização de fibra óptica para suprir as demandas de serviços multimídia e derivados, tem-se um forte investimento em tecnologias que ainda utilizam cobre para suprir tais demandas e é neste cenário que o trabalho desta dissertação se encaixa.

Até a presente data, não se tem conhecimento de trabalhos que se preocupem com a qualificação de enlaces considerando um grupo de PTs, ao invés disso, o que se tem é o foco de trabalhos que fazem qualificação de enlaces individuais. A qualificação de um conjunto de PTs, i.e., um *binder*, é um passo importante para os novos sistemas baseados em cobre em que as linhas são coordenadas. Novas tecnologias de cobre e sistemas modernos carregam consigo restrições de funcionamento que perpassam por (entre outras): distância máxima aceitável entre ponto de distribuição e cliente, e gerenciamento de espectro entre os PTs que se encontram no mesmo *binder* e oferecem os mais variados serviços xDSL. Algumas das principais discussões são apresentados a seguir e servem como motivação

para este trabalho.

As tecnologias modernas como *Femtocells*, pequenas estações rádio base (ERB) desenvolvidas para operar dentro de residências, que utilizam conexão banda larga existente na própria residência para se conectar à rede da operadora, operam em uma faixa de frequência que se sobrepõem a tecnologias xDSL existentes. Como mostrado na Figura 3 e descrito no trabalho (1), a tecnologia *FemtoWoC* possui restrições com relação à distância que o serviço pode operar sem que haja alta degradação do sinal. Esta degradação do sinal ocorre devido a interferências que serviços xDSL legado proporcionam ao novo serviço, haja vista que há uma sobreposição de largura de banda. Quanto maior for o compartilhamento entre os PTs no mesmo cabo, maior é a degradação do sinal do novo serviço. Por isso, um importante estudo sobre o comprimento de cabo d (Figura 4, extraída do mesmo artigo) cuja as tecnologias (novas e antigas) compartilham é um assunto abordado pelo trabalho (1). No contexto da estimação de comprimento de *binders*, a descoberta do valor de distância d auxilia na tomada de decisão a cerca do tipo de serviço que pode (ou não) ser oferecido à linha de transmissão que chega até o cliente.

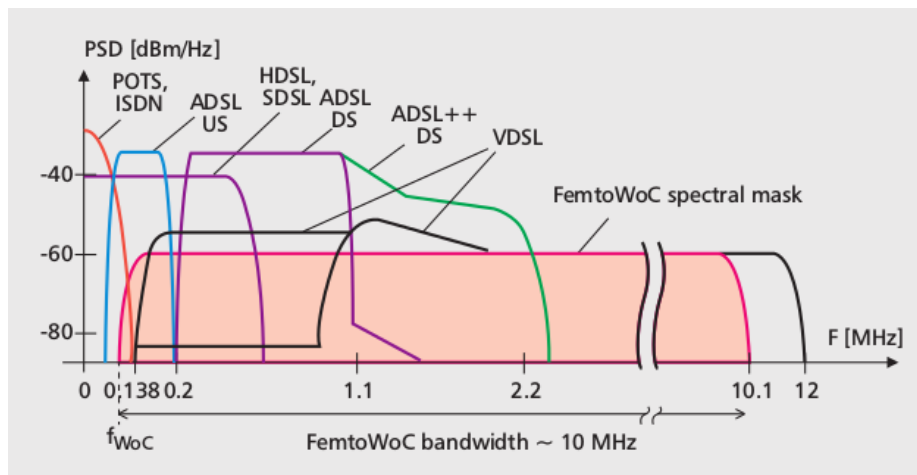


Figura 3 – Mapa espectral de serviços xDSL em linhas de telefone e coexistência com a tecnologia *FemtoWoC*. Extraído de: (1).

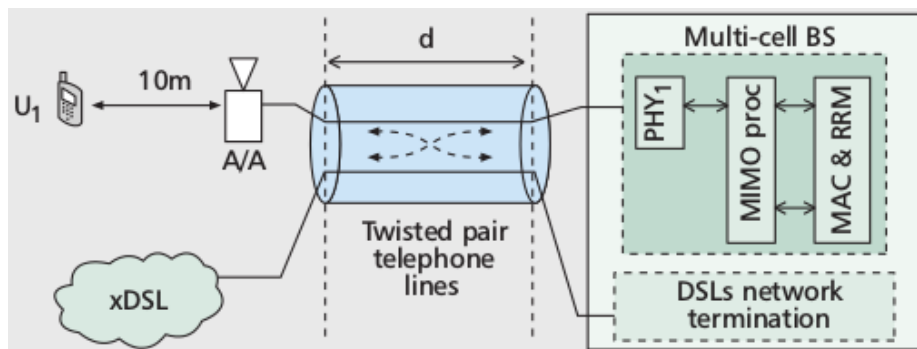


Figura 4 – *FemtoWoC* sujeito a interferência de outras linhas no mesmo *binder* trafegando serviços xDSL. Adaptado e extraído de: (1).

Outra questão importante a se destacar é a evolução de tecnologias xDSL ao longo do tempo. Atualmente, fala-se em 5ª geração de banda larga com a tecnologia *XG.fast*, onde promete-se taxas de transmissão agregada de até 10 Gb/s a uma distância de 70 metros. Testes em laboratório comprovam a eficácia desta tecnologia (2, 9). Assim como *G.fast*, o *XG.fast* é uma interessante alternativa para cobrir a última milha em redes banda larga. Na Figura 5, mostra-se a relação fibra-cobre juntamente com a tecnologia usada na infraestrutura POTS e o comprimento de cobre usado até a casa do cliente.

Admitindo que uma infraestrutura *fibra até a casa* (com sigla FTTH do inglês, *Fiber To The Home*) ainda não é uma solução factível em todos os lugares, e que usar a infraestrutura de cobre existente (POTS) é atrativa, pois – entre outras vantagens – já está implantada e o custo de manutenção é menor se comparado a fibra, sistemas híbridos fibra-cobre tornam-se soluções interessantes para ainda manter altas taxas de transmissão, quando da transição entre os dois meios físicos. Ao longo do tempo, as tecnologias xDSL contribuem significativamente com a aproximação da fibra à casa do cliente, uma vez que para se obter altas taxas de transmissão em cobre, mantendo níveis aceitáveis de ruído, é necessário diminuir o comprimento de enlace entre ponto de acesso e cliente. As principais tecnologias e suas respectivas distâncias são mostradas na Figura 5.

Inevitavelmente, em um dado momento no futuro ter-se-á um sistema totalmente coberto por fibra. A Figura 5 anuncia esta afirmação. Atualmente, já é possível vislumbrar esta realidade em alguns lugares do mundo, e.g., Emirados Árabes Unidos, onde FTTH atende 85% das residências (34). Entretanto, em localidades onde esta arquitetura ainda está longe de acontecer, as tecnologias xDSL implantadas em sistemas híbridos têm um papel importante para conectar a última milha ao cliente e, ainda, garantem baixo custo de manutenção e implantação.

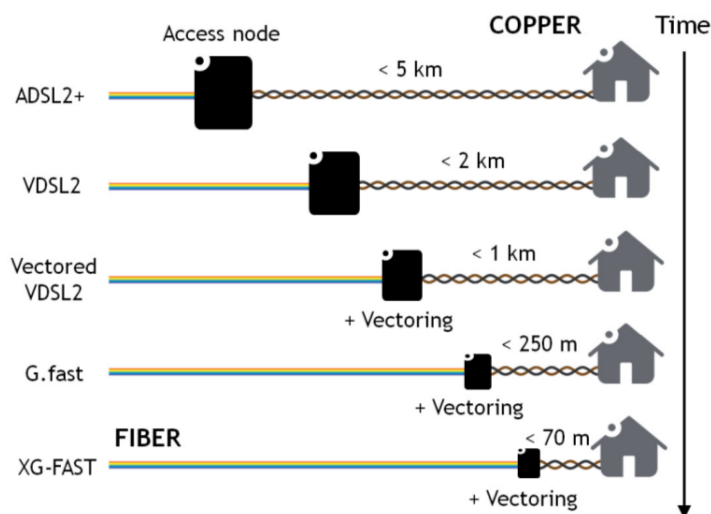


Figura 5 – Evolução de tecnologias de banda larga em fios de cobre e a relação de distância entre o ponto de distribuição e a casa do cliente em sistemas híbrido fibra-cobre. Extraído de: (2).

Outro trabalho mais recente viabilizou a utilização conjunta de *G.fast* e transmissão em modo fantasma em cenários comuns limitados por *crossstalk* do tipo FEXT (do inglês, *far-end crossstalk*) (31). O trabalho traz a discussão sobre benefícios que o modo de transmissão fantasma pode proporcionar na prática. Em suma, o trabalho mostrou que o desempenho do *G.fast* conjuntamente com modulação discreta em multiton e técnicas de *vectoring* para eliminar ruído (inclusive no canal fantasma) melhoram o desempenho de transmissão do *G.fast*, alcançando taxas agregadas acima de 2 Gb/s.

Visto que tecnologias xDSL, mitigação de ruído *crossstalk*, transmissão em modo fantasma, sistemas híbridos de fibra-cobre estão em voga na literatura, esta dissertação contribui para o avanço em sistema de transmissão de dados baseados em cobre através do desenvolvimento de um sistema automático de identificação de *binder*, possibilitando identificar a distância que os pares trançados compartilham no mesmo *binder* ou cabo. Este trabalho colabora com o avanço no desenvolvimento desses novos sistemas, tornando-os mais gerenciais, dinâmicos, factíveis e menos susceptíveis a ruídos e faltas. Este trabalho contribui na expansão do conceito de qualificação de um enlace para a qualificação de *binders* e cabos, além de servir de ferramenta, assim como uma técnica de detecção de falta, para identificar quais pares trançados de uma rede de *binders* podem ser utilizados para implantar essas novas tecnologias.

1.3 Justificativa

O trabalho descrito em (15) evidenciou a possibilidade de identificação de *binder* utilizando medições fantasmas. No trabalho em questão, a autora explica que realizar identificação de *binder* observando similaridades no domínio do tempo entre medições individuais de PTs em modo diferencial pode comprometer os resultados, pois nem sempre PTs que se encontram no mesmo *binder* terão comportamento semelhante ao analisar as medidas no domínio do tempo e da frequência, devido à não homogeneidades inerentes de qualquer par trançado. O mesmo ocorre para medições de PTs que estão em *binders* diferentes, i.e., o comportamento do sinal pode conter similaridades que apontem a mesma localidade dos PTs, i.e., apontem para PTs situados no mesmo *binder* .

Um método de identificação de *binder* que seja baseado em similaridades de respostas individuais de PTs requer uma *expertise* capaz de distinguir essas características individuais dos pares trançados. A proposta da autora é que ao invés de analisar similaridades entre respostas individuais de PTs, as similaridades sejam analisadas através de medições em modo fantasma que poderá definir a origem dos pares trançados.

A contribuição da autora é apresentar em seu trabalho uma prova de conceito de que é possível saber se o canal fantasma formado por dois PTs caracteriza-os como PTs que estão no mesmo *binder* ou em *binders* diferentes, e ainda, é feita uma discussão sobre a

possibilidade de estimar o comprimento do *binder* em que os PTs compartilham. Entretanto, o trabalho da autora resumiu-se a apresentar os desafios de realizar a identificação de *binders*, incluindo as provas de conceito, mas tão pouco se preocupou em desenvolver um algoritmo para aplicar na prática tal prova. Além disso, observou-se através das medições e descrições apresentadas no trabalho da autora, que na verdade estuda-se identificação de cabo e não de *binder*.

Assim, esta dissertação vem para preencher uma lacuna presente na literatura, propondo a identificação automática de *binders*. Além de verificar se pares trançados estão no mesmo cabo, também identifica se pares trançados estão dentro do mesmo *binder*. Uma metodologia para identificação de *binders* é proposta e testada através de algoritmos de aprendizado de máquina capazes de minerar bases de dados em busca de padrões. Os algoritmos de aprendizado de máquina são testados e avaliados em uma base de dados contendo medições fantasmas obtidas de um *setup* construído em laboratório.

Uma outra contribuição deste trabalho é a investigação rigorosa realizada em todos os atributos que serão extraídos das medições fantasmas, tanto no domínio do tempo quanto no domínio da frequência. A escolha desses atributos é fundamental para que os algoritmos de aprendizado de máquina possam trabalhar e gerar resultados aceitáveis. Ainda, é proposto um método de rotulação de *clusters* para realizar a classificação automática das amostras com os algoritmos de clusterização. Além disso, este trabalho também contribui com um método proposto que estima o comprimento em que os PTs compartilham o mesmo *binder*.

1.4 Objetivo Geral

Este trabalho tem como objetivo geral realizar a identificação automática de *binders* através de medições de pares trançados que estão transmitindo em modo fantasma. Pretende-se descobrir se o sinal em modo fantasma é oriundo de pares trançados que se encontram no mesmo *binder*, *binders* diferentes (e no mesmo cabo) ou em cabos diferentes. Para atingir este objetivo, um estudo rigoroso dos atributos extraídos das medições fantasmas será realizado. Uma base de dados contendo todas as medições fantasmas coletadas em laboratório é utilizada. As medições fantasmas serão obtidas através de um *setup* construído em laboratório. Os algoritmos de aprendizado de máquina serão aplicados na base de dados afim de identificar cada um dos padrões (classes) deste problema.

1.5 Objetivos Específicos

Cada objetivo específico constitui-se também de contribuições deste trabalho. Uma breve explicação sobre cada objetivo específico é dada a seguir.

- Elaborar um *setup* em laboratório e realizar medições em modo fantasma de pares trançados que se encontram em três ambientes distintos: mesmo *binder*, *binders* diferentes e cabos diferentes. Cada ambiente corresponde a um padrão (classe) que os algoritmos de aprendizado de máquina terão que identificar. Ressalta-se que cada medição fantasma é tratada para que seja representada por atributos. Os algoritmos de aprendizado de máquina trabalham nos atributos de cada medição fantasma. Este procedimento de extração de atributos será apresentado posteriormente;
- Determinação de atributos a partir de medições em modo fantasma. Dentre todos os atributos extraídos de cada medição fantasma, um estudo é apresentado com o objetivo de selecionar a melhor combinação de atributos para realizar o reconhecimento de padrões;
- Aplicar algoritmos de aprendizado de máquina para detectar padrões na base de dados com o objetivo de classificar cada amostra como **mesmo *binder*, *binders* diferentes** ou **cabos diferentes**. Os algoritmos de aprendizado de máquina que são explorados neste trabalho são: Máquina de Vetores de Suporte (com sigla SVM do inglês, *Support Vector Machine*), *K*-means e Modelo de Misturas de Gaussianas (com sigla GMM do inglês, *Gaussian Mixture Model*);
- Propor e testar um algoritmo para a estimação do comprimento em que os enlaces seguem ao longo do mesmo *binder* ou cabo. Esta estimação, obviamente, só poderá ser realizada caso uma medição fantasma for classificada como de mesmo *binder* ou de *binders* diferentes, i.e., medições de PTs situados no mesmo cabo;
- Incorporar automaticamente o conhecimento de rotulação de *clusters* (interpretados como padrões) aos modelos produzidos pelos algoritmos de clusterização (*K*-means e GMM) para torná-los aptos à classificação das medições fantasmas.
- Também é objetivo deste trabalho investigar a natureza do problema quanto ao aprendizado supervisionado e não supervisionado, disponibilizando as duas abordagens para criação de um modelo. O SVM será o algoritmo de aprendizado de máquina supervisionado. Já o *K*-means e GMM serão os algoritmos não supervisionados deste trabalho.

1.6 Estrutura da Dissertação

Esta dissertação está organizada da seguinte forma: o Capítulo 2 trata da transmissão em modo fantasma e também explicará os principais elementos que serão abordados nesta dissertação, juntamente com a qualificação de enlace. O Capítulo 3 apresenta os conceitos sobre os algoritmos de aprendizado de máquina e validação cruzada que são trabalhados nesta dissertação e a relação dos algoritmos na identificação de *binders*. Já

no Capítulo 4, mostra-se o algoritmo proposto para a identificação automática de *binder* e cabo, a análise de características, rotulação de *clusters* e a estimação de comprimento de compartilhamento. Os resultados dessa dissertação são apresentados no Capítulo 5. O trabalho finaliza com a conclusão e as publicações desta dissertação.

2 Transmissão em Modo Fantasma

Neste capítulo serão apresentadas definições de termos importantes deste trabalho que serão amplamente utilizados em nos próximos capítulos. Inicialmente é feita a contextualização dos elementos físicos principais que serão utilizados para realizar medições em modo fantasma. Posteriormente são feitas considerações sobre os principais modos de sinalização. Conclui-se o capítulo com a denominação do problema da identificação de *binder* como sendo uma ferramenta que pode ser usada para qualificar enlaces.

2.1 Terminologias e Contextualização do Trabalho

2.1.1 Par Trançado de Cobre

Par trançado (PT) é um cabo formado por um par de fios entrelaçados um ao redor do outro. O trançado é usado para cancelar a interferência eletromagnética de fontes externas e interferências mútuas (linha cruzada ou, em inglês, *crosstalk*). Para realizar a comunicação entre dois dispositivos, através de par trançado, é transmitida a mesma corrente elétrica nos dois fios, mas em direções opostas. Todos os cabos que foram usados em testes neste trabalho são formados por dois pares trançados. Esta configuração é uma característica importante para a formação dos circuitos fantasmas que serão vistos em seções posteriores. Outra característica importante é que pares trançados da rede de telefonia são agrupados em *binders*.

2.1.2 *Binder*

Define-se *binder* como sendo um conjunto de pares trançados de cobre que são trançados conjuntamente ao longo do caminho entre a central e o último ponto de distribuição próximo ao cliente. Geralmente 20 pares trançados ou mais podem ser agrupados em um mesmo *binder* (35). Eventualmente, pares trançados de um mesmo *binder* podem ser divididos em dois ou mais outros *binders* ao longo deste caminho, ramificando-se e formando a topologia da rede de transmissão de dados. Um cabo é constituído por um ou mais *binders*.

2.1.3 Cabo

Um cabo aglomera um conjunto de *binders*. Neste trabalho, quando se refere a pares trançados que estão no **mesmo cabo**, tem-se dois cenários: pares trançados no mesmo *binder* e pares trançados em *binders* diferentes, mas ainda no mesmo cabo. Enquanto

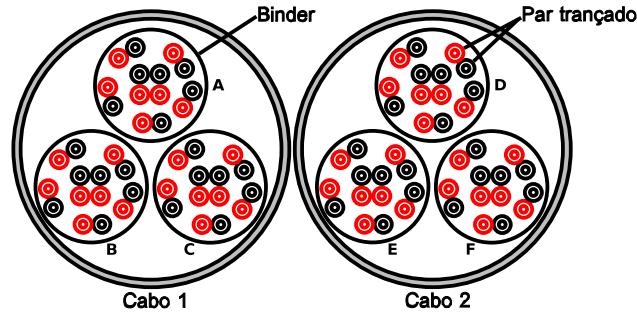


Figura 6 – Visão transversal de dois cabos contendo três *binders* cada um. Cada *binder* contém sete pares trançados.

que pares trançados em **cabos diferentes**, sabe-se que obviamente estão em *binders* diferentes.

A Figura 6 mostra uma seção transversal de dois cabos telefônicos. Doravante, toda vez que este trabalho se referir aos pares trançados que estão no mesmo *binder*, *binders* diferentes ou cabos diferentes será considerada a definição de cada termo a seguir. Considerando a Figura 6, formalmente temos que $C_1 = \{A, B, C\}$, $C_2 = \{D, E, F\}$.

- Mesmo *binder*: dois pares trançados $p_1, p_2 \in \theta$, tal que $\theta \in C_k, \forall k \in \{1, 2\}$, i.e., enquadram-se nesta definição quaisquer dois pares trançados que se encontram unicamente no mesmo *binder*;
- *Binders* diferentes: dois pares trançados $p_1 \in \theta_1$ e $p_2 \in \theta_2$, $\theta_1 \neq \theta_2$, tal que $\theta_1, \theta_2 \in C_k, \forall k \in \{1, 2\}$, i.e., considera-se desta definição quaisquer dois pares trançados que se encontram em *binders* diferentes, e.g., um par trançado do *binder A* e outro par trançado do *binder C*, ou quaisquer outras combinações de *binders* do **Cabo 1**. A mesma definição é válida para o **Cabo 2**, mas nunca uma combinação de pares trançados do **Cabo 1** e **2**.
- Cabos diferentes: dois pares trançados $p_1 \in \theta_1$ e $p_2 \in \theta_2$, tal que $\theta_1 \in C_1$ e $\theta_2 \in C_2$, e obviamente $\theta_1 \neq \theta_2$. Esta definição admite apenas pares trançados que se encontram em cabos diferentes, e.g., um par trançado situado no *binder B* do **Cabo 1** e outro par trançado situado no *binder F* do **Cabo 2**, e quaisquer outras combinações desta mesma forma.

Portanto, facilmente é possível generalizar as definições acima para uma situação onde se tem N_c cabos, cada cabo contendo N_b *binders*, e cada *binder* formado por N_p pares trançados. As definições acima são válidas e corretas para quaisquer tuplas $\{N_c, N_b, N_p\}$.

2.2 Modo de Sinalização

As sinalizações são formas de enviar um sinal em um circuito. Dentre elas, temos as formas mais comuns chamadas modo comum e modo diferencial. Entretanto, também se pode usar a transmissão em modo fantasma (20, 24). Este modo permite o uso de um canal virtual para enviar sinais entre o transmissor e o receptor. Por ser o modo de transmissão usado neste trabalho, o modo fantasma será mais detalhado nesta seção. Inicialmente, são apresentados o modo de sinalização comum e diferencial.

2.2.1 Modo Comum

A forma mais simples de sinalização é o modo comum onde um ou mais condutores são utilizados para o envio de um sinal. A Figura 7 ilustra um circuito com transmissor (Tx) e receptor (Rx) ao qual dois condutores são utilizados para enviar o sinal. Os potenciais elétricos que saem de Tx (V_1 e V_2) têm a mesma magnitude e sentido. O sinal recebido no Rx é calculado através da média aritmética entre todas os potenciais recebidos.

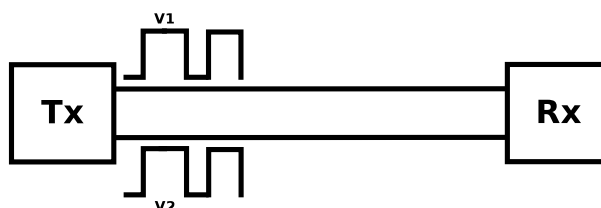


Figura 7 – Circuito com sinalização em modo comum.

Um dos problemas clássicos desse modo de sinalização é a indução eletromagnética. A interferência ocorre quando o condutor se encontra próximo de uma entidade que gera campo eletromagnético capaz de alterar o fluxo de corrente que passa pelos condutores. Esse problema pode ser contornado através do modo de sinalização diferencial.

2.2.2 Modo Diferencial

O modo diferencial também utiliza dois condutores para realizar a sinalização. No entanto, o potencial elétrico ocorre em sentidos contrários, i.e., um condutor tem potencial elétrico positivo e o outro condutor tem potencial elétrico negativo, como mostra a Figura 8. O módulo dos potenciais elétricos deve ter o mesmo valor.

O sinal é calculado no Rx através da diferença de potencial de sinais nos dois condutores. A consequência disso é uma maior robustez quanto à interferências em modo comum, como mostrado na Figura 9, pois em sistemas ideais a indução eletromagnética atinge da mesma forma os dois condutores e isto faz com que ela se anule no cálculo do sinal no Rx.

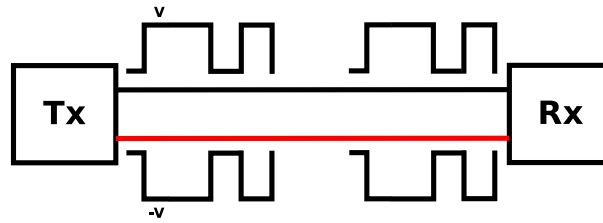


Figura 8 – Circuito com sinalização em modo diferencial.

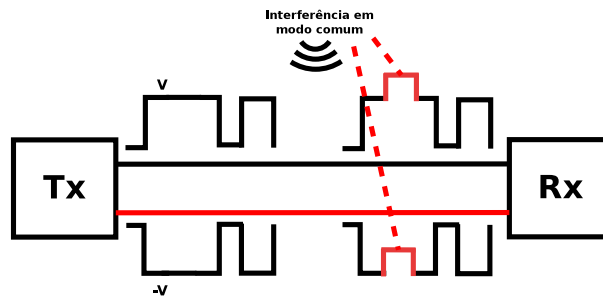


Figura 9 – Interferência em modo comum em um sistema com sinalização em modo diferencial.

2.3 Transmissão em Modo Fantasma

O modo de sinalização diferencial ajuda a mitigar problemas relacionados à interferências e conseqüentemente melhora a taxa de transmissão no par trançado de cobre. Já a transmissão em modo fantasma, além do circuito físico, é possível ter um circuito virtual usando mais de um par trançado, aumentando assim a taxa total de transmissão no par trançado. Enquanto que no modo diferencial tem-se a possibilidade de transmitir N sinais (onde N é o número de pares trançados), em modo fantasma, se houver restrição de que um par trançado só pode participar de apenas um canal fantasma, tem-se o número total de $\frac{3N}{2}$ sinais (N obrigatoriamente deve ser um número par), onde $\frac{N}{2}$ são sinais fantasmas de primeira camada (25).

O sistema de telefonia utiliza-se de transmissão em modo diferencial nas linhas telefônicas e é comum que cheguem mais de um par trançado nas instalações do cliente (geralmente dois pares trançados, mas apenas um par é utilizado para transmissão de sinal). O circuito fantasma permite um melhor uso da infraestrutura da rede de telefonia, visto que utiliza os dois pares trançados que já chegam à maioria dos clientes para formar canais adicionais de transmissão, aumentando assim o número de canais para transmissão de sinais entre a central e o cliente. Para utilizar esses canais virtuais, não há necessidade de operadores visitarem a casa do cliente, uma vez que os procedimentos necessários para se utilizar de circuitos fantasmas são feitos apenas no lado da central. Apenas medições do tipo SELT (do inglês, *Single Ended Loop Test*) são utilizadas para averiguar a viabilidade de implementação de circuitos fantasmas na linha telefônica.

A Figura 10 ilustra o funcionamento de um circuito fantasma através de dois pares

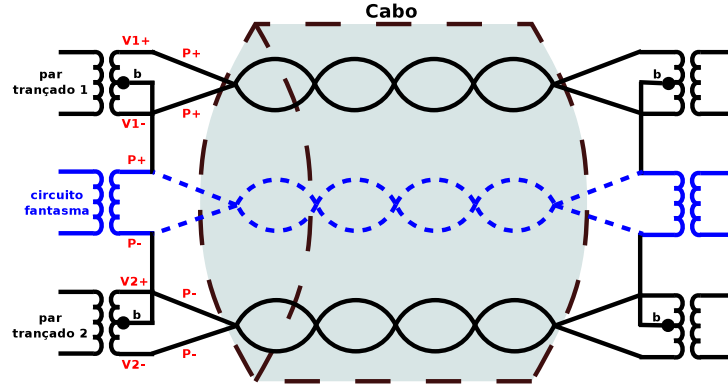


Figura 10 – Circuito fantasma utilizando dois pares trançados. Os *baluns* **b** são transformadores utilizados para fazer com que parte do sinal fantasma trafegue no modo comum de cada par trançado.

trançados. Os dois pares trançados estão transmitindo em modo diferencial representados $V1$ e $V2$. No modo fantasma, cada par trançado é visto como apenas um fio, e isso possibilita fazer com que um par trançado sirva como par de referência ou par de retorno de sinal (na Figura 10 o par de referência é o último). Este mecanismo faz com que o modo fantasma tenha características de um modo diferencial, trafegando simultaneamente em quatro fios através de sinais em modo comum (24), aqui representados por P .

O sinal no modo fantasma é calculado através da diferença entre as médias aritméticas nos dois condutores de cada par, i.e., levando em consideração a notação da Figura 11, tem-se $P+$ e $P-$ referentes à sinais em modo comum, enquanto que $V1$ e $V2$ são sinais transmitidos em modo diferencial, portanto o sinal fantasma pode ser calculado através da Equação 2.3 abaixo, sujeitando-se a

$$\begin{aligned}
 S_1 &= \frac{(V1+) + (P+) + (V1-) + (P+)}{2} \\
 &= \frac{(V1+) + (P+) + (-V1+) + (P+)}{2} \\
 &= \frac{2P+}{2} \\
 &= P+
 \end{aligned} \tag{2.1}$$

$$\begin{aligned}
 S_2 &= \frac{(V2+) + (P-) + (V2-) + (P-)}{2} \\
 &= \frac{(V2+) + (P-) + (-V2+) + (P-)}{2} \\
 &= \frac{2P-}{2} \\
 &= P-
 \end{aligned} \tag{2.2}$$

$$R = S_1 - S_2 \tag{2.3}$$

Em um sistema perfeitamente balanceado, os sinais em modo comum e diferencial não interferem entre si. Também não há interferência entre modo diferencial e fantasma, entretanto na prática é possível perceber interferências entre os modos.

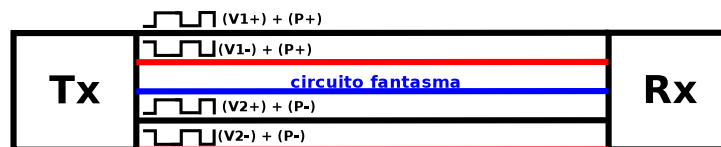


Figura 11 – Circuito com transmissão em modo fantasma.

2.4 Qualificação de Enlace

A linha telefônica foi originalmente desenvolvida para trabalhar a uma frequência de 4 kHz, enquanto sistemas DSL usam frequências mais altas. Diversos fatores devem ser levados em consideração antes de implantar um sistema DSL em uma linha telefônica: qualidade do cabo entre central e cliente, mudança do tipo de cabo e bitola ao longo da rede de telefonia, *bridged taps*, entre outros (16). Todos esses fatores são estudados e analisados para melhorar ou estender o uso do serviço de telefonia básica (em inglês, *Plain Old Telephone Service*). No entanto, uma decisão equivocada na escolha da combinação de fatores pode comprometer a qualidade do sinal oferecido por um serviço DSL. Nem todas as linhas são apropriadas para trabalhar com redes DSL. A qualificação de enlace é todo mecanismo utilizado para verificar a máxima taxa de transmissão que uma linha telefônica pode oferecer através de um sistema DSL e consequentemente definir o tipo de serviço que pode ser oferecido ao cliente.

A identificação de *binder* pode atuar como uma técnica de qualificação de enlace que ajuda a estimar o comprimento em que pares trançados compartilham o mesmo cabo, ou seja, através da estimação desse comprimento é possível identificar o local onde começam as instalações do cliente. Essa característica é primordial para realizar a qualificação do enlace, pois saber a distância entre o ponto de distribuição e a casa do cliente pode definir o nível de atenuação que o sinal sofre ao longo da linha, além de servir como parâmetro de estudo para o cenário de ruído entre central e cliente. Esse recurso também é importante, pois auxilia na identificação de faltas que podem ocorrer na linha de transmissão, onde a operadora de telefonia pode determinar se o problema ocorreu nas dependências de suas instalações ou nas dependências do cliente, e assim, reduzindo custos na investigação manual que os operadores realizariam.

2.5 Identificação de Topologia do Enlace

As técnicas de identificação de topologia consistem em obter informações a respeito do enlace, tais como: comprimento total do enlace, número de seções do enlace, comprimento e bitola de cada seção. Estas informações são importantes, pois auxiliam na tomada de decisão a cerca de qual enlace deve ser selecionado para prover um sistema DSL.

Há pelo menos três métodos para se qualificar um enlace telefônico: (i) extraíndo informações de um banco de dados que contenha as plantas telefônicas de uma determinada área; (ii) através de medições DELT (*Double Ended Loop Test*); e (iii) SELT (*Single Ended Loop Test*). As plantas telefônicas eram historicamente mantidas em papéis e posteriormente foram sendo introduzidas em banco de dados. Entretanto, a inclusão manual das plantas gerou um problema quanto à precisão dos dados encontrados nos bancos de dados e inconsistência quanto à atualização desses dados (devido a expansão da rede de telefonia), tornando este método de qualificar enlace quase impraticável (17). O segundo método, DELT, permite facilmente uma análise do enlace, porém há a necessidade de enviar um técnico às dependências do cliente para instalar equipamentos usados para se comunicar com outros equipamentos que se encontram na central. Já o terceiro método, SELT, requer apenas um equipamento que se encontra na central. Este método consome menos tempo e custo em relação ao método DELT, visto que não há necessidade de enviar um técnico às dependências do cliente.

Os métodos de identificação de topologia da literatura são predominantemente baseados em medições SELT, devido às dificuldades de se obter medições de duas portas DELT. As medições SELT são analisadas através da reflectometria no domínio do tempo (36) ou no domínio da frequência (37). O método desenvolvido neste trabalho utiliza medições SELT, com finalidade de analisar as reflexões do sinal no domínio do tempo e descobrir o momento em que os pares trançados se separam, i.e., deixam de compartilhar o mesmo cabo.

Saber a distância de compartilhamento é importante em diversas aplicações DSL, pois via de regra define o desempenho do sistema em função da distância entre a central e o cliente. Em tecnologias mais recentes, como o *FemtoWoC*, saber a distância em que pares trançados compartilham a mesma estrutura impacta diretamente nos recursos que podem ser disponibilizados pela tecnologia (1). Em sistemas DSLs mais recentes, como *G.fast* (5) e *XG.fast* (9, 2), apresentam limitações quanto à distância máxima que essas tecnologias conseguem alcançar sem que ocorram perdas significativas na qualidade do sinal, operando em 1 Gb/s a uma distância de 250 m e 10 Gb/s a 70 m. Portanto, o algoritmo proposto para identificação de *binders* e a descoberta do comprimento de compartilhamento em que os pares trançados compartilham o mesmo *binder*, auxiliam na escolha de quais linhas podem ser oferecidas tais serviços, além disso outras contribuições agregadas nessas informações adquiridas pelos algoritmos podem auxiliar no: melhor gerenciamento da

linha de transmissão, estudo de *Crosstalk* em determinados *binders* , desenvolvimento e implantação de tecnologias futuras que trabalham em modo híbrido com fibra-cobre.

3 Aprendizado de Máquina na Identificação de *Binders*

Este capítulo contempla a teoria dos algoritmos de aprendizado de máquina que são abordados nesta dissertação. Estes algoritmos são utilizados para alcançar e concluir os objetivos deste trabalho. A Máquina de Vetores de Suporte, é um algoritmo utilizado em sua forma de aprendizado supervisionada, i.e., cada amostra da base de dados necessita ter previamente sua classificação disponível, pois este tipo de aprendizado utiliza esta classificação em suas formulações matemáticas com o intuito de reter o conhecimento de padrões na base de dados. Os algoritmos K -means e Modelo de Misturas de Gaussianas, trabalham com o aprendizado não supervisionado, i.e., não necessitam que as amostras da base de dados tenham sua respectiva classificação disponível.

Os algoritmos de aprendizado de máquina são utilizados em diversas tarefas de mineração de dados, aos quais se encontram: classificação, predição, regras de associação, clusterização e detecção de *outlier*. Em linhas gerais, um modelo pode ser produzido através de uma base de dados $\mathbf{B} \in \mathbb{R}^{m \times n}$, com m amostras e n atributos. A base \mathbf{B} é dividida em duas outras bases: a) $\mathbf{T} \in \mathbb{R}^{p \times n}$, onde \mathbf{T} é a base de treino, p é o número de amostras contidas nesta base; b) $\mathbf{Z} \in \mathbb{R}^{m-p \times n}$, onde \mathbf{Z} é a base de teste. O algoritmo de aprendizado de máquina escolhido deve ser capaz de produzir um modelo a partir da base \mathbf{T} e validá-lo na base \mathbf{Z} . A simbologia apresentada neste parágrafo será usada ao longo deste capítulo.

3.1 Máquina de Vetores de Suporte

É uma técnica de aprendizado de máquina não paramétrica supervisionada amplamente utilizada em tarefas de mineração de dado, e.g., classificação de dados e regressão (38, 39). Por ser um algoritmo de aprendizado supervisionado, necessita que previamente todas as amostras da base de treinamento já estejam previamente classificadas. As amostras da base de treinamento são usadas para construir um modelo que contém informações necessárias para classificar amostras que não foram utilizadas no processo de treinamento.

Em sua forma mais simples, o SVM constrói modelos binários, i.e., quando uma amostra é apresentada a um modelo SVM, este se encarrega de classificar a amostra através de uma função sinal seguinte

$$\text{sign}(f(x)) = \begin{cases} -1, & f(x) < 0 \\ 0, & f(x) = 0 \\ 1, & f(x) > 0 \end{cases}, \quad (3.1)$$

onde x é uma amostra que se deseja classificar, $f(\cdot)$ é a função/modelo do SVM que contém todos os parâmetros específicos deste algoritmo para classificar uma amostra. Percebe-se pela Equação 3.1 que a saída de um modelo SVM é binária (1/ -1), a saída nula (0) é um caso especial no treinamento do modelo onde a amostra x se caracteriza como um vetor de suporte. Na saída binária cabe interpretação externa sobre o significado desta binarização. O que geralmente se faz é entender que todas as amostras cujo resultado for -1 são consideradas da mesma classe, e portanto possuem similaridades em comum. E todas as amostras classificadas como 1 são pertencentes a outra classe e também possuem similaridades em comum.

O algoritmo SVM tem como objetivo principal encontrar hiperplanos ótimos que maximizam a separação entre dois conjuntos de dados, i.e., entre as duas classes que formam a base de dados de treinamento, de tal forma que quando uma amostra for apresentada ao modelo, este a classificará como pertencente a um dos dois conjuntos. Por trabalhar com hiperplanos, o SVM divide os dois conjuntos de forma linear. A Figura 12 mostra um exemplo de hiperplano ótimo encontrado pelo algoritmo SVM, nela tem-se uma base de dados com dois conjuntos (quadrado e círculo) e um conjunto de classificadores lineares (retas) que dividem a base em duas classes, mas apenas um classificador (em destaque) separa a base, maximizando a margem entre as duas classes. Os vetores de suporte são calculados através do algoritmo SVM.

O exemplo mostrado na Figura 12 retrata uma base de dados onde há a separação linear das amostras em duas classes, porém muitas bases de dados que modelam problemas reais apresentam comportamento não linear. Nestes casos, o algoritmo SVM se utiliza de um recurso chamado de função kernel. Se considerarmos cada amostra como um vetor, a função kernel é responsável por mapear cada amostra da base de dados que se encontra em seu espaço característico original Ψ' com d_1 dimensões para um espaço característico Ψ'' de dimensão superior d_2 , onde $d_2 > d_1$. Neste espaço característico com d_2 dimensões o

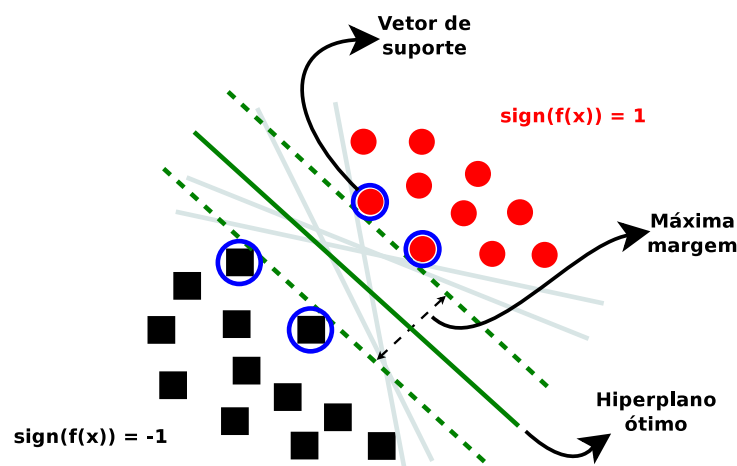


Figura 12 – Candidatos a hiperplanos de separação e hiperplano ótimo em destaque.

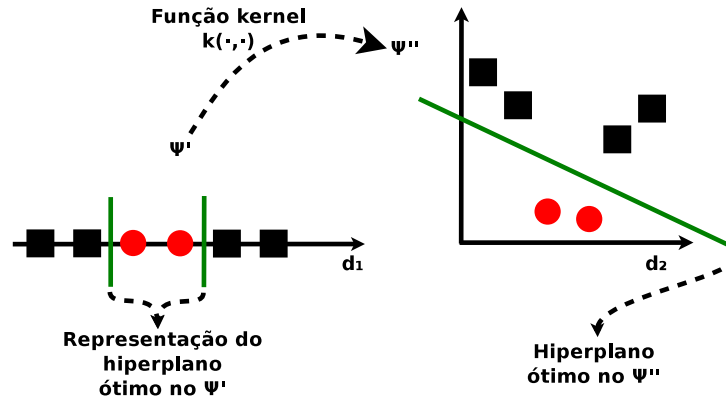


Figura 13 – Exemplo de uma função kernel $k(\cdot, \cdot)$ para separação de classes em problemas não lineares.

problema é tratado como linear, podendo então dividir a base de dados em dois conjuntos através do hiperplano. Um exemplo desse recurso do algoritmo é mostrado na Figura 13. Nesta figura, tem-se um espaço característico com $d_1 = 1$ dimensão cuja as amostras de duas classes (quadrado e círculo) não podem ser separadas linearmente. Através de uma função kernel é possível separá-las em dois conjuntos em um espaço característico com $d_2 = 2$ dimensões. Note que a representação do hiperplano ótimo em Ψ' é formada por apenas dois pontos.

Classificar uma amostra pode ser interpretado como descobrir o grau de similaridade que esta amostra tem com a base de dados de treinamento. Desta forma, a função kernel é vista como uma medida de similaridade que permite construir algoritmos no espaço Ψ'' (40). Em termos matemáticos, a função kernel consiste em realizar um produto interno entre duas amostras no espaço característico Ψ' . Existem várias funções kernel na literatura. Neste trabalho, apenas a função polinomial será destacada, pois é esta função que será usada para construir o modelo SVM na base de dados de treinamento. A função kernel polinomial é expressada pela equação seguinte

$$k(x, x_i) = (x^T \cdot x_i + c)^{|x_i|+1}, \quad (3.2)$$

onde $k(\cdot, \cdot)$ é a função kernel polinomial, x e x_i são amostras do espaço Ψ' , c é uma constante e $|\cdot|$ é a cardinalidade do vetor. O kernel polinomial é da ordem $|x_1| + 1$. A função de decisão é definida na equação seguinte

$$f(x) = \sum_{i=1}^m \alpha_i y_i k(x, x_i) + b, \quad (3.3)$$

onde α_i e y_i são, respectivamente, multiplicador Lagrangeano e a classe associada ao vetor de suporte x_i (amostra da base de dados de treinamento), m é o número de vetores de suporte e b um parâmetro otimizado da função linear. Esta função pode resultar em um valor positivo ou negativo, a classificação de uma nova amostra x é feita através da resolução da função sinal apresentada na Equação 3.1 já apresentada.

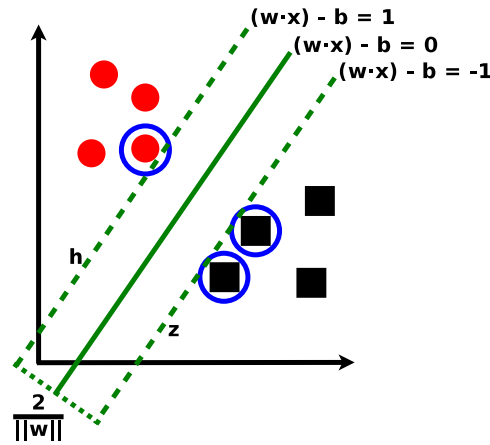


Figura 14 – Exemplo de um hiperplano com margem maximizadas em uma base de dados com duas classes. As amostras circuladas são vetores de suporte.

Encontrar um hiperplano de máxima margem para um base de treino é um problema de otimização. Tomando como exemplo a Figura 14, é possível selecionar dois hiperplanos de uma maneira que não haverá amostras entre eles e tentar maximizar a distância que os separa através da fórmula

$$d_{h,z} = \frac{2}{\|w\|}. \quad (3.4)$$

Para maximizar $d_{h,z}$, deve-se minimizar $\|w\|$. Algumas restrições importantes devem ser consideradas para que nenhuma amostra recaia entre as margens. São elas:

$$\text{restrições} = \begin{cases} w \cdot x_i - b \geq 1 & \forall x_i \mid y_i = 1 \\ w \cdot x_i - b \leq -1 & \forall x_i \mid y_i = -1 \end{cases}.$$

Essas restrições podem ser reescritas considerando toda a base de treino

$$y_i(w \cdot x_i - b) \geq 1 \quad \forall x_i \mid y_i \in \{-1, 1\}. \quad (3.5)$$

Na forma primal, o algoritmo SVM precisa minimizar $\|w\|$. Por conveniência matemática, pode-se substituir $\|w\|$ por $\frac{1}{2}\|w\|^2$ sem alterar a solução. Este é considerado um problema de otimização quadrática da forma

$$\arg \min_{(w,b)} \frac{1}{2} \|w\|^2, \quad (3.6)$$

sujeito à Equação 3.5. Através da introdução dos Multiplicadores de Lagrange α , a restrição anterior pode ser expressa da forma

$$\arg \min_{(w,b)} \max_{\alpha \geq 0} \left\{ \frac{1}{2} \|w\|^2 - \sum_{i=1}^n \alpha_i [y_i(w \cdot x_i - b) - 1] \right\}, \quad (3.7)$$

onde $w = \sum_{i=1}^n \alpha_i y_i x_i$, com restrição de $\sum_{i=1}^n \alpha_i y_i = 0$. Apenas alguns α_i serão maiores que 0. As amostras x_i correspondentes serão os vetores de suporte que recaem sobre a margem e satisfazem a restrição mostrada na Equação 3.5.

3.2 K-means

É um algoritmo de aprendizado de máquina não supervisionado que realiza a tarefa de clusterização. Clusterizar é o processo de particionar ou agrupar um conjunto de amostras em *clusters* disjuntos. Amostras de um mesmo *cluster* são similares e amostras de *cluster* distintos são dissimilares.

O algoritmo utilizado neste trabalho é aquele descrito em (41). O número de *clusters* K é definido previamente e é fixo. Cada *cluster* é representado pelo posicionamento de seu centroide (w_1, \dots, w_K) . Cada $w \in \mathbb{R}^n$ é um vetor que possui n dimensões e é inicializado em alguma das p amostras da base \mathbf{T} definidas por (i_1, \dots, i_p) . Então,

$$w_j = i_l, \quad j \in \{1, \dots, K\}, l \in \{1, \dots, p\}.$$

Cada dimensão $d \in \{1, \dots, n\}$ de um centroide w_j é calculado através da média dos valores das amostras que pertencem ao *cluster* j na respectiva dimensão d . A qualidade da clusterização é determinada pela minimização do erro na equação seguinte

$$E = \sum_{k=1}^K \sum_{i_l \in C_j} \|i_l - w_k\|^2. \quad (3.8)$$

O algoritmo K -means funciona de forma iterativa descrito sucintamente nas etapas seguintes:

1. Calcula a distância de cada amostra p para cada centroide w_j ;
2. Atribui cada amostra p a um *cluster* j mais próximo;
3. Recalcula a posição dos centroides w através da média dos valores das amostras pertencentes ao *cluster* em cada dimensão d .
4. Calcula E e avalia a condição de parada. O algoritmo para quando não há redução do valor E ;
5. Se não satisfazer a condição de parada, volta para a etapa 1.

Especificamente nas etapas 1 e 2, deve-se calcular a distância para cada centroide e definir a qual *cluster* cada amostra p pertence, respectivamente. Para realizar estas etapas uma métrica de distância deve ser escolhida. Dentre as diversas métricas de distância disponíveis na literatura, este trabalho optou pela distância euclidiana quadrática definida a seguir

$$D_i = \sum_{d=1}^n \|i_{l,d} - w_{j,d}\|^2, \quad (3.9)$$

onde D_i é a distância da amostra i_l para o centroide w_j do *cluster* que ela pertence.

É importante ressaltar o significado de cada *cluster* encontrado na base de dados de treinamento. Os *clusters* representam regiões com alta densidade, sendo que cada um representa uma classe do problema, i.e., cada *cluster* está diretamente relacionado a um padrão da base e, portanto, deve ter significado revelante para os especialistas, pois são essas regiões de alta densidade que podem ser efetivamente rotuladas por eles. Rotular um *cluster* significa atribuir uma classe a ele, e assim, tornar possível calcular a acurácia de um modelo que realiza clusterização. Este trabalho também apresenta um algoritmo para realizar esta tarefa, detalhes sobre sua implementação serão apresentados na Seção 4.5.

Outras características importantes da tarefa de clusterização são que os centroides de cada *cluster* devem estar razoavelmente longe um do outro e a média de distância entre as amostras e o centroide mais próximo deve ser mínima. Se a base possuir informação suficiente, o mínimo encontrado pelo algoritmo tende a ser próximo ou igual ao mínimo global. Figura 15 mostra um exemplo do algoritmo *K*-means em um espaço bidimensional definido por três *clusters*. É possível observar que o modelo produzido por este algoritmo é definido apenas pelas posições dos centroides, no exemplo em questão, tem-se $K = 3$ representando o centro de massa de cada aglomeração de amostras.

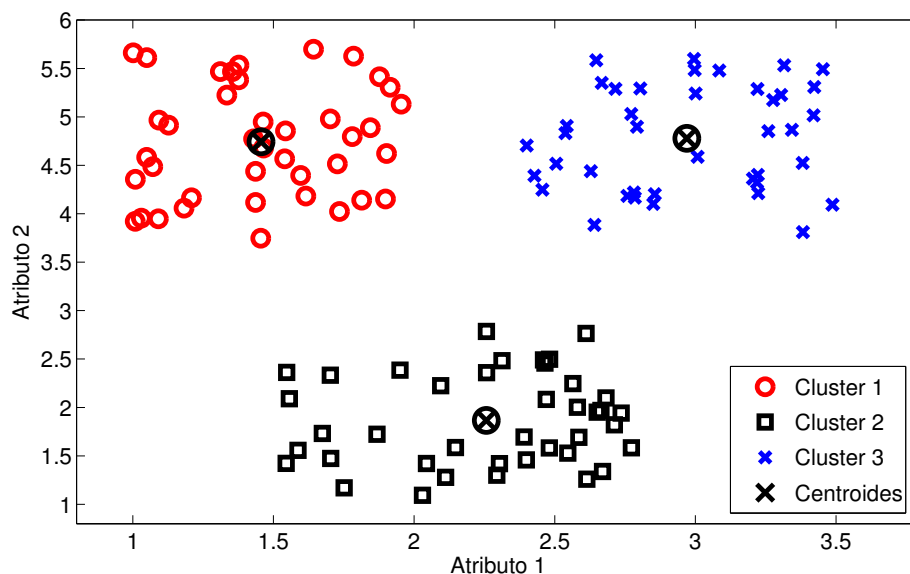


Figura 15 – Exemplo do resultado de uma clusterização utilizando o algoritmo *K*-means.

3.3 Modelo de Misturas de Gaussianas

Este algoritmo constrói um modelo probabilístico cujo objetivo é estimar os principais componentes de Gaussianas em uma base de dados de treinamento. Os componentes principais de uma Gaussiana j são formados pela tupla $(\mu_j, \Sigma_j, \alpha_j)$, onde μ_j é a média da Gaussiana, Σ_j a covariância e α_j um parâmetro de proporção, onde $\sum_{k=1}^K \alpha_k = 1$. A

probabilidade de uma amostra x_i pertencer a um componente (μ_j, Σ_j) está sujeito a função densidade de probabilidade

$$f(x_i | \mu_j, \Sigma_j) = \frac{1}{\sqrt{(2\pi)^m |\Sigma_j|}} e^{-1/2(x_i - \mu_j)^T \Sigma_j^{-1} (x_i - \mu_j)}. \quad (3.10)$$

Neste trabalho, cada Gaussiana é interpretada como um *cluster*, o que implica dizer que pela simbologia adotada pelo algoritmo K -means, tem-se $\mu_j = w_j$. GMM é um técnica de clusterização suave, o que significa que um *cluster* pode sobrepor outro *cluster*, i.e., uma Gaussiana pode se posicionar próxima o suficiente de outra Gaussiana, sobrepondo-se. E uma amostra não pertence necessariamente a um *cluster* apenas (como era o caso do algoritmo K -means). Agora cada amostra recebe uma probabilidade de pertencer a um determinado *cluster* chamada de Probabilidade Posterior (Equação 3.11), trazendo mais flexibilidade aos resultados do algoritmo.

$$r_{i,j} = \frac{\alpha_j f(x_i | \mu_j, \Sigma_j)}{\sum_{k=1}^K \alpha_k f(x_i | \mu_k, \Sigma_k)}. \quad (3.11)$$

Os parâmetros $(\mu_j, \Sigma_j, \alpha_j)$ são estimados através do algoritmo de Expectativa / Maximização (com sigla EM do inglês, *Expectation Maximization*) (42), sendo que na primeira iteração a tupla é inicializada aleatoriamente. Esse algoritmo executa de forma iterativa em dois passos:

1. Passo E: calculam-se as probabilidades das amostras da base de treino pertencer a cada *cluster* (Equação 3.11);
2. Passo M: atualizam-se os parâmetros $(\mu_j, \Sigma_j, \alpha_j)$ de cada *cluster* através das Equações 3.12, 3.13, 3.14, respectivamente.

$$\mu_{j,d} = \frac{1}{p\alpha_j} \sum_{i=1}^p r_{i,j} x_{i,d} \quad (3.12)$$

$$(\Sigma_j)_{t,k} = \frac{1}{p\alpha_j} \sum_{i=1}^p r_{i,j} (x_{i,t} - \mu_{j,t})(x_{i,k} - \mu_{j,k}) \quad (3.13)$$

$$\alpha_j = \frac{1}{p} \sum_{i=1}^p r_{i,j} \quad (3.14)$$

A mistura de todas as distribuições gaussianas é formulada através *log-likelihood* (Equação 3.15). O algoritmo GMM termina após um determinado número de repetições ou até que o *log-likelihood* tenha convergido para um ótimo local.

$$L = \log L(T) = \sum_{i=1}^p \log \sum_{k=1}^K \alpha_k f(x_i | \mu_k, \Sigma_k) \quad (3.15)$$

3.4 SVM, K -means e GMM na Identificação de *Binders*

As duas abordagens de aprendizagem, supervisionada e não supervisionada, são utilizadas nesta dissertação para extrair padrões de bases de dados que tenham relação com os três cenários possíveis considerados: mesmo *binder*, *binders* diferentes e cabos diferentes. A abordagem supervisionada requer que todas as amostras coletadas e armazenadas na base de dados estejam com suas classes previamente definidas. Os algoritmos de aprendizado de máquina que utilizam abordagem supervisionada geralmente fazem uso desta informação para construir modelos de classificação. O modelo armazena todas as informações necessárias para classificar novas amostras, i.e., amostras que não foram apresentadas ao algoritmo na etapa de treinamento. A abordagem não supervisionada não requer que as amostras da base de dados estejam classificadas. Algoritmos que usam essa abordagem geralmente trabalham com medidas de similaridades entre as amostras, fazendo com que amostras que são similares entre si sejam classificadas como de um mesmo grupo, e amostras diferentes entre si sejam classificadas como de grupos diferentes.

No contexto do problema da identificação de *binders*, que é o objetivo principal deste trabalho, tem-se uma situação onde um operador terá a oportunidade de obter medições em modo fantasma de pares trançados que se encontram em um determinado cenário, e.g., a rede de *binders* de uma cidade. Na ocasião, o objetivo do operador é exatamente descobrir a origem desses pares trançados, i.e., se dois pares se encontram no mesmo *binder*, *binders* diferentes ou em cabos diferentes. Então, o que se percebe aqui é que o operador pode automatizar este processo de descoberta sendo orientado por um classificador. No entanto, para construir esse classificador que automatiza o processo de classificação de uma medição fantasma, o operador terá que construir uma base de dados contendo medições fantasmas do mesmo ambiente, onde as amostras não terão a sua classe previamente definida, pois é exatamente esta informação que o operador deseja obter. Neste momento, tem-se a caracterização de que o operador terá que utilizar um algoritmo de aprendizado de máquina cuja abordagem seja não supervisionada, pois as amostras coletadas não terão classes predefinidas. Por esse motivo, este trabalho opta por testar a proposta de algoritmo de identificação de *binders* através do K -means e GMM.

O algoritmo SVM, que utiliza abordagem supervisionada, também é utilizado para efeito de comparação. E também pode ser uma opção para o operador caso o mesmo disponha de um laboratório devidamente equipado com todo o aparato tecnológico necessário para construir um *setup* de medições em modo fantasma, numa tentativa de simular o cenário da mesma rede de *binders* ao qual o operador deseja realizar as classificações. Nesta ocasião, o operador poderá construir um modelo mais especializado e mais robusto, haja vista que a base de dados que se poderá construir conterá as classes das medições fantasmas feitas em laboratório. Como alternativa, a validação de um classificador construído desta forma pode ser feita com medições fantasmas do cenário real.

Os algoritmos K -means e GMM têm como parâmetro comum a variável k . Esta variável tem como objetivo modelar o número de *clusters* para o K -means e o número de componentes (Gaussianas) para o GMM. Neste sentido, sabe-se previamente que o valor da variável $k = 3$ para o problema da identificação de *binders*, pois são três padrões que deseja-se extrair da base de dados de treinamento. É importante destacar nesta dissertação que a presença de amostras *outliers* não é considerada, i.e., todas as medições em modo fantasma recaem em características semelhantes a um dos três cenários avaliados.

Um ajuste particular deve ser feito para utilizar o algoritmo SVM no problema da identificação de *binders*, pois o problema requer o reconhecimento de três padrões, mas o algoritmo SVM trabalha de forma binária com saídas $\{-1, 1\}$. O ajuste consiste em converter o problema da identificação de *binders* em um problema binário, admitindo uma nova classe chamada *mesmo cabo* = $\{ \text{mesmo binder} \cup \text{binders diferentes} \}$ e a outra classe *cabos diferentes*. Este ajuste é feito na base de dados \mathbf{B} . Após este procedimento, o algoritmo SVM é executado normalmente conforme descrito na Seção 3.1. Este recurso permitirá criar um classificador π_1 . Para completar o problema da identificação de *binders*, deve-se distinguir também as amostras que pertencerem a nova classe *mesmo cabo*. Para realizar esta tarefa, um novo classificador π_2 deve ser construído utilizando a base de dados $\mathbf{B} = \mathbf{B} - \text{cabos diferentes}$. Com isso, tem-se a base de dados \mathbf{B} contendo apenas amostras *mesmo binders* e *binders diferentes*, ou seja, recaindo em um problema binário e, portanto, pode-se executar o algoritmo SVM normalmente. De posse dos dois classificadores π_1 e π_2 , o procedimento para classificar uma amostra x qualquer é dado pelo Pseudocódigo 1.

Algoritmo 1: Classificação de uma amostra com SVM.

Entrada: x, π_1, π_2
Saída: Classe da amostra x

```

1 início
2    $r := null$ ;
3   se  $\pi_1(x) = \text{mesmo cabo}$  então
4     se  $\pi_2(x) = \text{mesmo binder}$  então
5        $r := \text{mesmo binder}$ 
6     senão
7        $r := \text{binders diferentes}$ 
8     fim
9   senão
10     $r := \text{cabos diferentes}$ 
11  fim
12 fim
13 retorna  $r$ 

```

Primeiramente é feita uma classificação com o classificador π_1 (linha 3) para saber se a amostra x pertence a classe *mesmo cabo*. Caso negativo, imediatamente a amostra é classificada como de *cabos diferentes* (linha 10). Caso positivo é necessário realizar a

segunda classificação com o classificador π_2 (linha 4) para descobrir se a amostra é oriunda de um mesmo *binder*. Caso positivo, a amostra é classificada como de *mesmo binder* (linha 5). E caso negativo, a amostra é classificada como de *binders diferentes* (linha 7). A forma de avaliação de um resultado do algoritmo SVM é feita através deste pseudocódigo, comparando a classe r que o algoritmo informou com a classe real da amostra.

3.5 Validação Cruzada Estratificada

Na mineração de dados uma das tarefas mais importantes do processo de aprendizagem é a avaliação dos modelos, principalmente quando se deve comparar resultados de algoritmos de aprendizado de máquina diferentes. Em linhas gerais, o procedimento para avaliar um modelo construído por um algoritmo pode ser dividido sucintamente em três passos.

1. Divide-se a base de dados \mathbf{B} em duas bases: treino \mathbf{T} e teste \mathbf{Z} ;
2. Executa o algoritmo na base de treino \mathbf{T} ;
3. Avalia o modelo construído pelo algoritmo na base de teste \mathbf{Z} ;

Um dos problemas relacionados a esses passos é quanto ao processo de validação do modelo (passo 3), principalmente se a base de dados \mathbf{B} contiver poucas amostras. Este tipo de problema se enquadra na tarefa de identificação de *binders*. Além disso, a acurácia de um modelo deve ser calculada em uma base (de teste) que contenha amostras que não foram usadas para treinar o modelo, i.e., as amostras da base de treino não devem ser usadas para testar o modelo. Isto é um fator importante quando se deseja apurar um modelo através de sua acurácia, sendo que a acurácia que o modelo produz na base de treino não deve ser considerada um indicador bom para generalizar a performance do modelo para futuras amostras, e quaisquer avaliações neste sentido serão consideradas otimistas (43).

Por esta razão, este trabalho executa várias vezes a validação cruzada na mesma base de dados e, desta forma, garante que não haverá resultados com valores elevados de acurácia alcançados através de um *efeito de chance* (geralmente atribuído aos algoritmos que possuem processos estocásticos), pois os modelos são avaliados através de médias estatísticas calculadas entre todos os modelos produzidos durante a validação cruzada. A explicação a seguir demonstra o procedimento da validação cruzada.

A validação cruzada é dita estratificada quando a base de dados é particionada aleatoriamente em k subconjuntos de mesmo tamanho (também chamado de $k - folds$) e há um esforço para manter a mesma proporção original de cada classe em cada *fold*. A validação cruzada estratificada executa k vezes, onde em cada execução um único *fold* é usada como um conjunto de teste e os remanescentes $k - 1$ *folds* são usados como conjunto

de treinamento. Cada *fold* é usado exatamente uma única vez como conjunto de teste. A Figura 16 demonstra um exemplo de validação cruzada com $k = 4$. Neste exemplo, o círculo representa uma base de dados que se dividiu em diferentes partes a cada execução. O *fold* hachurado mostra a porção da base destinada à validação do modelo π produzido em cada uma das execuções.

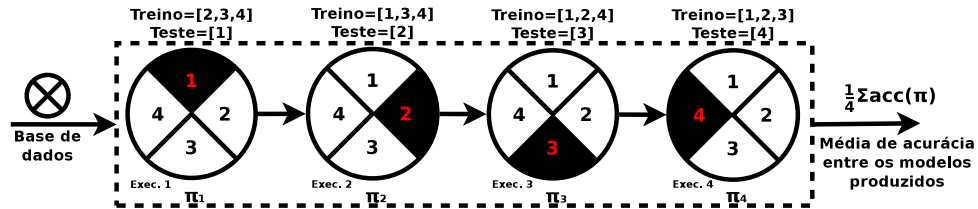


Figura 16 – Exemplo de uma validação cruzada com $k = 4$ aplicada a uma base de dados.

O objetivo da validação cruzada é generalizar os resultados através de análise estatística que represente intuições sobre como um modelo se comportará quando submetido a uma base de dados desconhecida. A validação cruzada facilita a identificação de *overfitting*, i.e., a especialização de um determinado modelo a sua base de treino. A saída é a média de acurácia de amostras classificadas corretamente em cada execução.

A base de dados \mathbf{B} é estratificada em 10 – *folds*. Este valor é escolhido por ter sido extensivamente utilizado em diferentes bases de dados com diferentes técnicas de aprendizado de máquina e parece ser o número correto para se estimar erros (43). Será feita uma combinação de características para a aplicação de cada algoritmo com o objetivo de encontrar a melhor combinação para o problema da identificação de *binders*. Após a descoberta, essa melhor combinação de características será explorada na mesma base de dados \mathbf{B} sob experimentos específicos.

4 Método de Identificação de *Binder*

Este capítulo explicará a proposta para a identificação de *binder*. Serão apresentadas todas as características extraídas de um sinal S_{11}^{PM} da base de medições fantasmas \mathbf{A} coletadas em laboratório e usadas para construir a base de dados $\mathbf{B} \in \mathbb{R}^{m \times n}$, com m amostras e n dimensões (características do sinal). Este capítulo inicialmente discute características sobre o sinal fantasma e a motivação para a escolha das características a serem extraídas. Em seguida, é apresentado a forma de extração de cada característica. Depois, a apresentação do algoritmo de identificação de *binders*, o método de rotulação de *clusters* e o método para a estimação de comprimentos.

4.1 Considerações Sobre Características do Sinal Fantasma

Na Figura 17 são mostradas 45 medições fantasmas (obtidas aleatoriamente da base de dados) para os três cenários avaliados. Visualmente, nota-se que o sinal em modo fantasma tem comportamento bem característico em cada um dos casos. O descasamento de impedância entre o equipamento de medição e o circuito fantasma influenciam rigorosamente na quantidade de energia que será introduzida no canal fantasma.

Quando dois PTs estão no mesmo *binder*, tem-se que a impedância característica do circuito fantasma formado por eles é baixa, pois a distância entre os PTs é menor, o que contribui para que haja forte acoplamento entre os pares e baixa oposição do sinal que é injetado no canal. Por outro lado, o equipamento de medição também possui impedância característica baixa. Uma vez que se tem ambas as impedâncias características próximas uma da outra, sabe-se que se obtém um descasamento de impedância baixo. Isto, conseqüentemente, implica em um baixo efeito NER e, portanto, boa parte do sinal em modo fantasma será injetado no canal. Uma porção de energia do sinal em modo fantasma contribuí para que se forme ondas estacionárias com alta periodicidade no domínio da frequência (linhas cheias da Figura 17).

Quando os dois PTs que formam o circuito fantasma se encontram em cabos diferentes ou no mesmo cabo, mas em *binders* diferentes, a distância entre os pares faz com que a impedância característica do canal fantasma seja alta, distanciando-se da impedância característica do equipamento de medição e fazendo com que ocorra um forte efeito NER. Portanto, boa parte do sinal em modo fantasma será refletido na entrada do canal, em proporções maiores para o cenário cabos diferentes e em proporções menores para o cenário de *binders* diferentes. No domínio da frequência, a periodicidade do sinal é menos perceptível no cenário *binders* diferentes, e quase imperceptível no cenário cabos diferentes. Este comportamento é visto na figura através das linhas tracejadas, para o

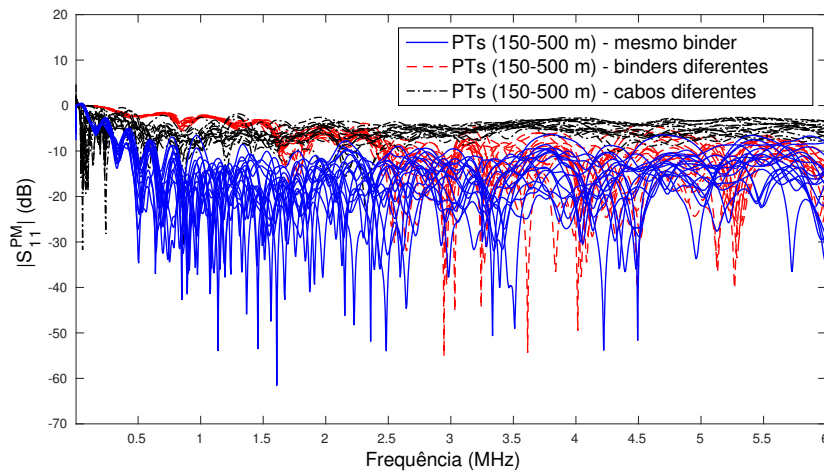


Figura 17 – Gráfico contendo 45 medições em modo fantasma. As curvas de com linhas no estilo ponto tracejadas representam as medições de PTs de diferentes cabos, as curvas com linhas no estilo tracejadas são medições de *binders* diferentes e as curvas de linhas cheias são medições de mesmo *binder*.

cenário de *binders* diferentes, e linhas traços e pontos para o cenário de cabos diferentes.

A influência do descasamento de impedância já foi observada e descrita no trabalho (15). Entretanto, apesar do referido trabalho apresentar medições de circuitos fantasmas de PTs no mesmo *binder* e em *binders* diferentes, acredita-se que os autores trabalharam com PTs situados no mesmo cabo e em cabos diferentes. Esta suposição justifica-se pelas medições apresentadas no artigo em questão. Além disso, não é discutido a influência da utilização de cabos blindados nas medições, o que pode ter levado a conclusão precipitada de que a identificação de *binders* é uma tarefa fácil. Mas em uma situação comum, onde a maioria dos PTs que compõem uma rede de telefonia não são blindados, i.e., sofrem interferências externas e há um “vazamento” de sinal para outros pares, a primeira característica afetada é a periodicidade do sinal que sofre alterações nos três cenários. Conseqüentemente, o sinal fantasma no domínio da frequência pode apresentar um comportamento que pode confundir os classificadores.

A Figura 18 mostra uma das situações que tornam o problema da identificação de *binders* uma tarefa difícil de ser realizada. Um circuito fantasma (CF 1) formado por dois PTs que estão no mesmo cabo, mas estão em *binders* diferentes. Outro circuito fantasma (CF 2) formado por dois PTs que estão no mesmo *binder*. No primeiro caso, CF 1 é composto por pares que estão próximos um do outro, porém em *binders* diferentes. Devido a proximidade e trançamento dos *binders*, os PTs sofrem acoplamento e são factíveis de transmitirem o sinal fantasma, fazendo com que o sinal se comporte semelhante ao padrão de mesmo *binder*. No segundo caso, os PTs envolvidos estão mais distantes (se comparados ao primeiro caso), entretanto se situam no mesmo *binder*. Os PTs também sofrem acoplamento, porém a distância entre os pares influenciará no comportamento do

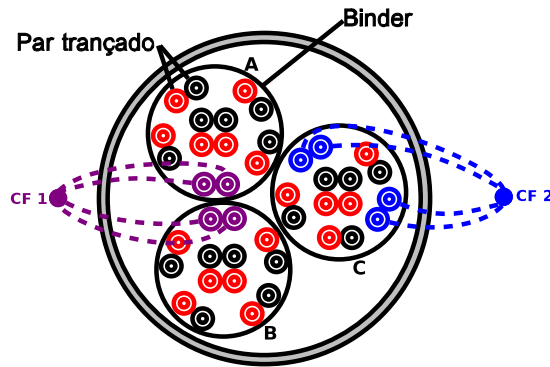


Figura 18 – Visão transversal de um cabo contendo três *binders* cada um. No exemplo, um circuito fantasma (CF 1) é formado entre dois PTs que estão em *binders* diferentes (A e B). Outro circuito fantasma (CF 2) é formado por dois PTs que estão no mesmo *binder* (C).

sinal no domínio da frequência, fazendo com que o sinal fantasma se assemelhe ao padrão de *binders* diferentes.

Em suma, ambos os sinais sofrem perda de periodicidade e apresentam um comportamento semelhante na forma de onda do sinal (veja um exemplo na Figura 19). Portanto, quaisquer estratégias usadas para extrair características deve levar em consideração esta situação. A Seção 4.4 apresenta e discute mais sobre a semelhança entre esses dois cenários, observando e analisando diretamente o espaço característicos das medições. Vale ressaltar que semelhante comportamento também pode ocorrer no cenário de cabos diferentes (admitindo que os cabos adjacentes não são blindados), mas este efeito será menos perceptível no domínio da frequência.

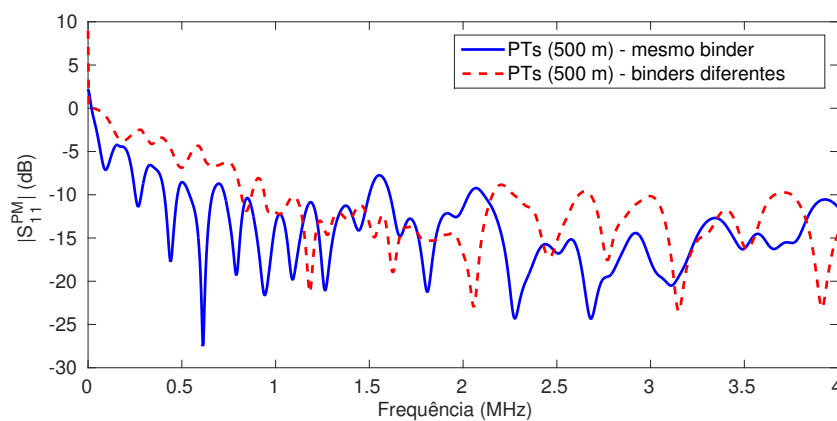


Figura 19 – Semelhança entre duas medições em modo fantasma oriundas de cenários diferentes. Em azul, dois PTs no mesmo *binder* e, em vermelho, dois PTs em *binders* diferentes.

4.2 Extração de Características do Sinal Fantasma

A extração de características do sinal fantasma consiste em uma transformação realizada na base \mathbf{A} para calcular atributos que representem cada sinal fantasma, na forma geral

$$A \xrightarrow{f(S_{11}^{PM})} B.$$

A base de medições \mathbf{A} contem todos os sinais fantasmas no domínio da frequência. A transformação é necessária para que os algoritmos de aprendizado de máquina trabalhem diretamente no espaço característico da base de dados \mathbf{B} , i.e., computacionalmente a entrada dos algoritmos é uma matriz contendo amostras dispostas em linha, onde cada uma armazena características extraídas de um sinal fantasma formado por PTs em um determinado cenário.

Para inserir os conceitos sobre cada característica que é extraída do sinal fantasma, deve-se primeiro estabelecer algumas notações listadas a seguir:

- $S_{11}^{PM} = \frac{V_{PM}^-}{V_{PM}^+} \rightarrow$ parâmetro de dispersão do sinal, onde V_{PM}^- e V_{PM}^+ são os sinais recebido e enviado, respectivamente (mais detalhado na Seção 5.1);
- $|S_{11}^{PM}| = 20 \times \log_{10}(\text{abs}(S_{11}^{PM})) \rightarrow$ amplitude do sinal usando escala logarítmica de decibel;
- $S_{11}^{PM}(f) \rightarrow$ sinal fantasma no domínio da frequência;
- $S_{11}^{PM}(t) \rightarrow$ sinal fantasma no domínio do tempo.

Os circuitos fantasmas formados por PTs que estão no mesmo *binder* apresentam características de um sinal modulado sofrendo atenuação. Assim, a presença de periodicidades no parâmetro S_{11}^{PM} revelou ser uma evidência para identificar PTs neste cenário (15). Neste sentido, duas características relacionadas à presença de periodicidade do sinal fantasma foram escolhidas: a variância do período no sinal $|S_{11}^{PM}|$ e análise das linhas espectrais de $S_{11}^{PM}(f)$, que consiste em aplicar a Densidade Espectral de Potência (com sigla PSD do inglês, *Power Spectral Density*) em $S_{11}^{PM}(f)$. Esta última característica é considerada uma transformação matemática para revelar a periodicidade do sinal. Duas outras características relacionadas ao efeito *Near-End Reflection* (NER) também são consideradas: a variância da magnitude do sinal $|S_{11}^{PM}|$ e o primeiro ponto do sinal no domínio do tempo $S_{11}^{PM}(t)$. Esta última está diretamente ligada ao nível de sinal que é refletido na entrada do canal e pode ser percebido no domínio do tempo. Este trabalho, doravante, denominará as características com a seguinte notação:

- f_1 , variância do período de $|S_{11}^{PM}|$: σ_p^2 ;

- f_2 , variância da magnitude de $|S_{11}^{PM}|$: σ_M^2 ;
- f_3 , número de linhas espectrais obtidos pela PSD em S_{11}^{PM} , $\mathcal{F}(S_{11}^{PM}(f))$: n_Φ ;
- f_4 , o primeiro ponto de reflexão da resposta do sinal no domínio do tempo calculado pela Transformada Inversa de Fourier do sinal S_{11}^{PM} , $\mathcal{T}(S_{11}^{PM}(f))$: \mathcal{A}_1 .

A variância do período (σ_p^2) é estimada através do cálculo da distância entre picos na frequência de $|S_{11}^{PM}|$. A Figura 20 apresenta um sinal em modo fantasma formado por dois PTs no mesmo *binder*. O sinal é dividido em n segmentos através de um espaçamento (*shift*) e para cada um desses segmentos é aplicada a equação seguinte

$$p_j = \frac{1}{n_j - 1} \sum_{i=1}^{n_j-1} \|t_{j,i+1} - t_{j,i}\|^2, \quad (4.1)$$

onde p_j é a variância do segmento j , $t_{j,i+1}$ e $t_{j,i}$ são posições de dois picos consecutivos no mesmo segmento j e n_j é o número total de picos em j . Finalmente, a média de todas as variâncias $E(P)$, onde $P \in \{p_1, p_2, \dots, p_n\}$, é a característica σ_p^2 . A medição S_{11}^{PM} de PTs no mesmo *binder* tem periodicidade bem definida, i.e., os comprimentos entre seções (representados por L_x , L_y e L_z na Figura 20) são aproximadamente iguais ao longo do domínio da frequência, e assim se mantém para quaisquer dois comprimentos entre dois picos consecutivos. Não se recomenda que seja utilizadas frequências acima de 4 MHz, pois a periodicidade do sinal torna-se menos evidente.

A variância da magnitude (σ_M^2) está mais relacionada com a influência causada pela atenuação do circuito fantasma no sinal $|S_{11}^{PM}|$. Dois PTs que estão em diferentes cabos formam um circuito fantasma com impedância característica elevada, causando forte efeito NER e, portanto, parte do sinal sofre reflexão logo na entrada do circuito

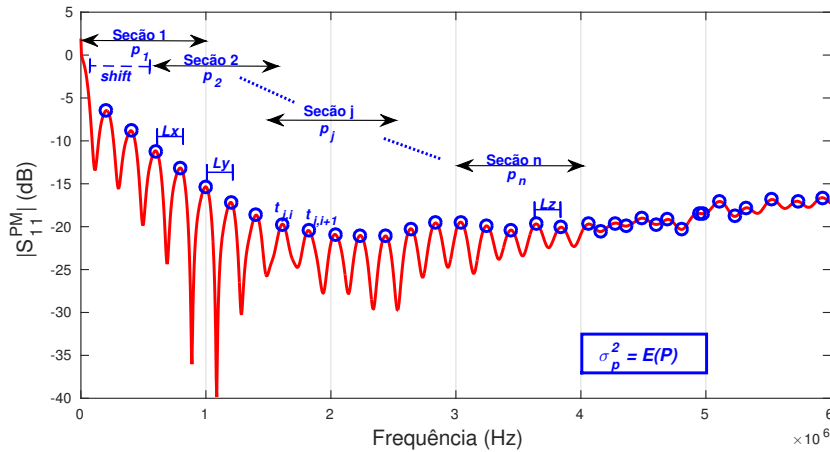


Figura 20 – Variância do período extraída no domínio da frequência de um sinal fantasma formado por dois PTs de 500 m de comprimento que estão situados no mesmo *binder*.

fantasma. O oposto ocorre quando dois PTs estão situados no mesmo cabo ou até mesmo no mesmo *binder*, a impedância característica do circuito fantasma é baixa e boa parte do sinal é propagado ao longo da linha, criando ondas estacionárias no domínio da frequência. Este efeito é intermediário quando dois PTs que formam o circuito fantasma se situam no mesmo cabo, mas em diferentes *binders*. A Figura 21 mostra um sinal fantasma com os picos devidamente detectados. Novamente, não se recomenda utilizar frequências acima de 4 MHz. O cálculo desta característica (Equação 4.2) é obtido através da variância da amplitude para todos os picos no sinal $|S_{11}^{PM}|$, conforme

$$\sigma_M^2 = \frac{1}{n-1} \sum_{j=1}^n pks_j - \overline{pks}, \quad (4.2)$$

onde \overline{pks} é a media de picos. Esta característica tenta capturar a diferença na amplitude do sinal $|S_{11}^{PM}|$ ao longo do domínio da frequência nos diferentes cenários de disposição dos pares trançados.

A terceira característica (n_Φ) tem relação com a primeira característica, porém utiliza-se de uma análise matemática para estimar a periodicidade do sinal. A PSD é um recurso aplicado ao sinal no domínio do tempo, este que é convertido para o domínio da frequência através da transformada de Fourier. Com o sinal no domínio da frequência, a densidade espectral de potência é aplicada, revelando os componentes harmônicos (chamados de linhas espectrais) do sinal (44). Nesta dissertação, esta característica aplica a PSD no $S_{11}^{PM}(f)$ como se fosse um sinal no domínio do tempo. Esta operação é indicada pela expressão $\mathcal{F}(S_{11}^{PM}(f))$. Depois de eliminar ruídos de frequências baixas com um filtro passa-altas, o método encontra o número de linhas espectrais em $\mathcal{F}(S_{11}^{PM}(f))$, chamado de n_Φ . Se n_Φ é baixo (< 10 picos, definido por inspeção), significa que $\mathcal{F}(S_{11}^{PM}(f))$ tem poucos picos, indicando que a periodicidade do sinal e a medição provavelmente correspondem a

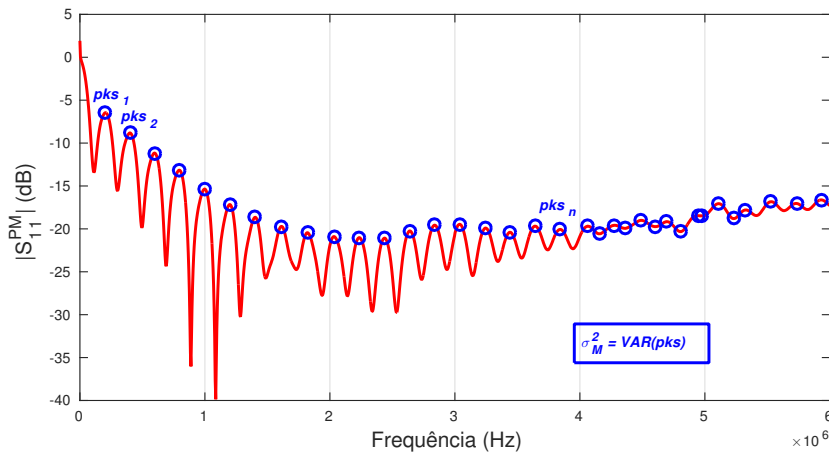


Figura 21 – Variância da magnitude extraída no domínio da frequência de um sinal fantasma formado por dois PTs de 500 m de comprimento que estão situados no mesmo *binder*.

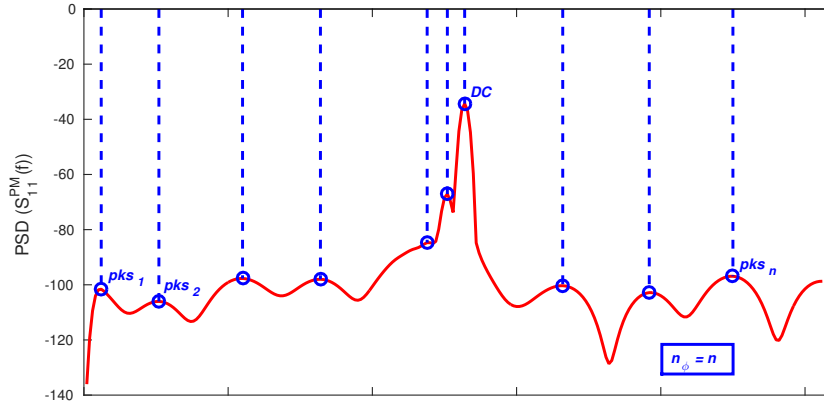


Figura 22 – Linhas espectrais aplicando PSD no domínio da frequência de uma medição em modo fantasma.

TPs no mesmo *binder*, caso contrário o cenário pode indicar PTs em *binders* diferentes ou, ainda, cabos diferentes. A Figura 22 apresenta um exemplo da PSD aplicada a um sinal $S_{11}^{PM}(f)$ de dois PTs no mesmo *binder*. No exemplo, $n_{\phi} = 10$, o que indica PTs no mesmo *binder*.

A quarta característica (\mathcal{A}_1) tenta capturar o nível de descasamento entre a impedância de entrada do circuito fantasma e a impedância do equipamento de medição, via NER. Através de experimentos de teste, pode-se reparar que esta característica se apresenta como uma forte assinatura do sinal fantasma no domínio do tempo, apresentando uma diferença clara entre os níveis de NER de PTs que estão no mesmo *binder*, *binders* diferentes e cabos diferentes (mais detalhado nas Seções 4.1 e 4.4). O método usa um operador \mathcal{T} que representa a reflectometria no domínio do tempo calculada através da transformada inversa de Fourier do S_{11}^{PM} , $\mathcal{T}(S_{11}^{PM}(f))$. A quarta característica é o primeiro ponto $\mathcal{T}(S_{11}^{PM}(f))$, chamado de \mathcal{A}_1 . Um exemplo desta característica é mostrado na Figura 23. Um sinal fantasma formado por dois PTs no mesmo *binder* é convertido para o domínio do tempo e, posteriormente, o primeiro ponto da curva TDR é capturado e considerado como quarta característica.

Uma vez que toda medição fantasma com diferentes pares trançados exibirá diferentes características, é possível caracterizá-la como uma amostra representada por um vetor $f = (\sigma_p^2, \sigma_M^2, n_{\phi}, \mathcal{A}_1) = (f_1, f_2, f_3, f_4)$, formando um espaço característico $\varphi \rightarrow \mathbb{R} \times \mathbb{R} \times \mathbb{Z} \times \mathbb{R}$. O resultado da extração de características de cada amostra da base de medições \mathbf{A} é o conjunto de amostras que formam a base de dados \mathbf{B} . Esta base de dados compreende o espaço característico φ , onde os algoritmos de aprendizado de máquina trabalham. O padrão de uma amostra é identificado baseando-se na sua localização neste espaço característico.

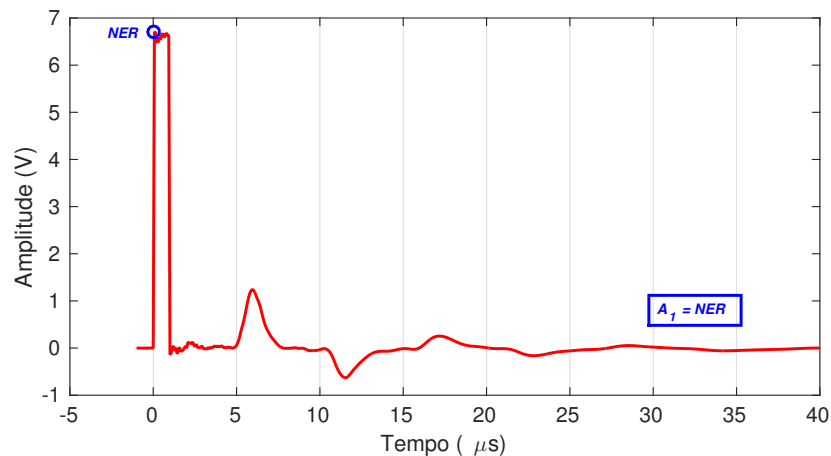


Figura 23 – Primeiro ponto da TDR de um sinal fantasma formado por dois PTs de 500 m de comprimento que estão situados no mesmo *binder*. A quarta característica tem relação direta com o primeiro ponto S_{11}^{PM} , que tem relação direta com efeito NER. O círculo representa o nível NER.

4.3 Algoritmo para a Identificação de *Binders*

O algoritmo proposto usa o reconhecimento de padrão para descobrir se uma medição S_{11}^{PM} é oriunda de dois PTs no mesmo *binder*. Para isto as características citadas na Seção 4.4 são utilizadas como informações representativas de uma medição. Os algoritmos de aprendizado de máquina trabalham diretamente nas características extraídas de uma medição S_{11}^{PM} .

Nesta proposta do algoritmo de identificação de *binders*, deve-se primeiramente pontuar algumas condições. Considere o problema da identificação de *binder*: saber se um PT^r (chamado PT de referência) está compartilhando o mesmo *binder* com outro PT^t (chamado PT de teste) e qual distância eles coexistem dentro deste mesmo *binder*. Duas condições são consideradas:

- Um PT^r contendo n seções (incluindo *bridged taps*). Se esse PT^r deixa de compartilhar o mesmo *binder* com o PT^t a partir de uma seção k_1 , tal que $k_1 = 1, 2, \dots, n - 1$, eles não compartilharão o mesmo *binder* em nenhuma seção posterior k_2 , tal que $k_1 < k_2 < n - 1$, nesta topologia.
- PTs com diferentes diâmetros de condutores nunca compartilham o mesmo *binder* (45, 46), pois considera-se que todos os PTs de um mesmo *binder* utilizam o mesmo diâmetro de condutor.

O método de identificação de *binder* é apresentado na Figura 24. Os três cenários possíveis de medição fantasma de dois PTs são assistidos nesta figura: *intra-binder*, *inter-binders* e *inter-cabos*.

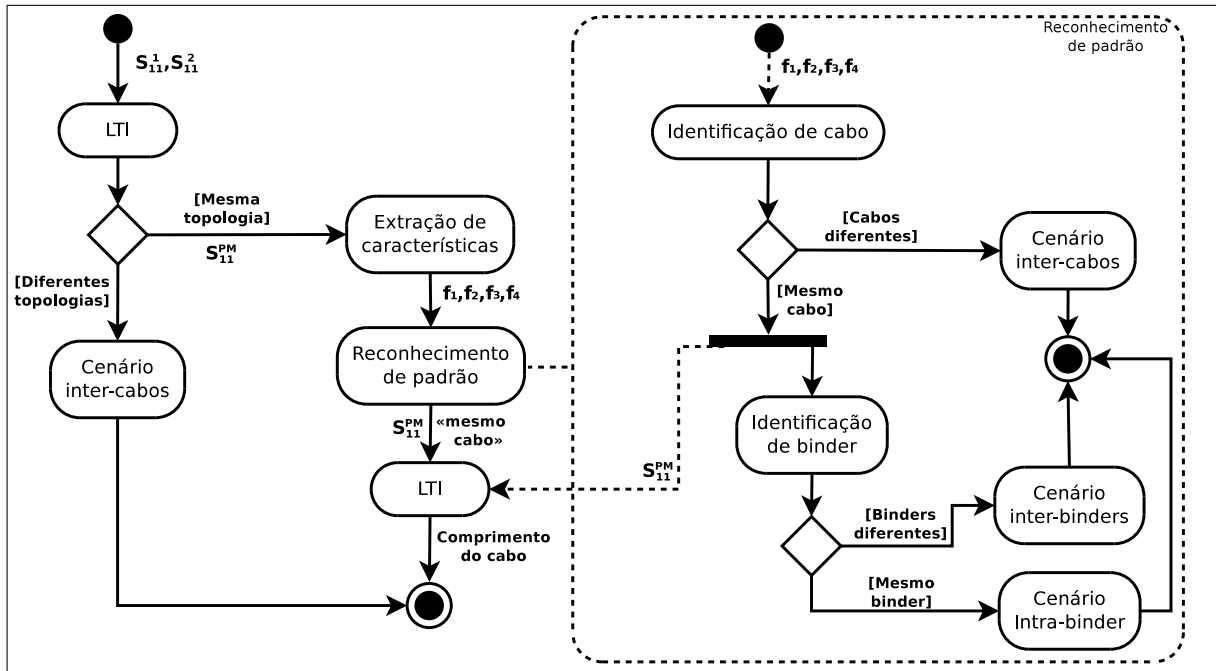


Figura 24 – Fluxograma do algoritmo da identificação de *binders*.

Esta figura funciona para um circuito fantasma (uma amostra) e também para um conjunto de medições (base de dados) S_{11}^{PM} , permitindo obter informações confiáveis sobre PTs de uma rede de distribuição. A seguir é comentado cada passo do fluxograma:

1. O PT de referência e o PT de teste são identificados usando uma técnica LTI para medições em modo diferencial;
2. Se os PTs têm resultados diferentes para a técnica LTI, em relação ao diâmetro do condutor, então eles estão no cenário inter-cabos;
3. Embora os resultados da técnica LTI possam ser diferentes em relação ao comprimento, os PTs podem estar em um cenário parcialmente compartilhado (15), i.e., eles podem estar compartilhando um mesmo *binder* em um trecho inicial e depois se dividirem em *binders* diferentes;
4. É feita a extração de características da medição S_{11}^{PM} ;
5. O passo de reconhecimento de padrão ocorre em subpassos responsáveis por classificar a medição S_{11}^{PM} . São eles:
 - a) Aplica a identificação de cabos nas características extraídas e devidamente selecionadas da medição S_{11}^{PM} ;
 - b) A classificação como cabos diferentes resulta no cenário inter-cabos, enquanto que a classificação de mesmo cabo requer uma inspeção adicional dentro do cabo;

- c) Essa inspeção, que é a última parte do reconhecimento de padrão, define dois cenários restantes possíveis: *binders* diferentes para inter-*binders* ou mesmo *binder* para intra-*binder*;
6. Após o passo de reconhecimento de padrão, aplica-se o último passo a técnica LTI na medição S_{11}^{PM} localizadas no mesmo cabo (independentemente dos PTs estarem no mesmo *binder* ou em *binders* diferentes) para descobrir a distância que os PTs coexistem no mesmo *binder*.

4.4 Análise das Características

Esta seção trata de analisar as características que são extraídas de um sinal fantasma. O estudo de características permite obter conhecimento do domínio do problema diretamente em seu espaço característico φ e, também, obtém-se um direcionamento para descobrir quais características são mais sensíveis para o problema da **identificação de *binders***. Dado que o algoritmo SVM constrói modelos binários, uma relação direta entre amostra de entrada e sua classificação propriamente dita passa por um passo intermediário cuja tarefa é dada através da **identificação de cabos**. A análise desta seção também leva em consideração o cenário em que as classes mesmo *binder* e *binders* diferentes são consideradas como de mesmo cabo. Através das Figuras 25, 26, 27 e 28, uma análise é feita para se entender a sensibilidade revelada pelas características.

A tupla $(\sigma_p^2, \sigma_M^2, n_\Phi, \mathcal{A}_1) = (f_1, f_2, f_3, f_4)$ caracteriza uma amostra no espaço φ através da função de extração de características $f(S_{11}^{PM})$. Considerando o problema da identificação de *binders* como sendo binário, tem-se duas classes para acomodar todas as amostras da base de dados: cabos diferentes e mesmo cabo (Figura 25). Esta última classe representa a junção das classes mesmo *binder* e *binders* diferentes. Nesta abordagem

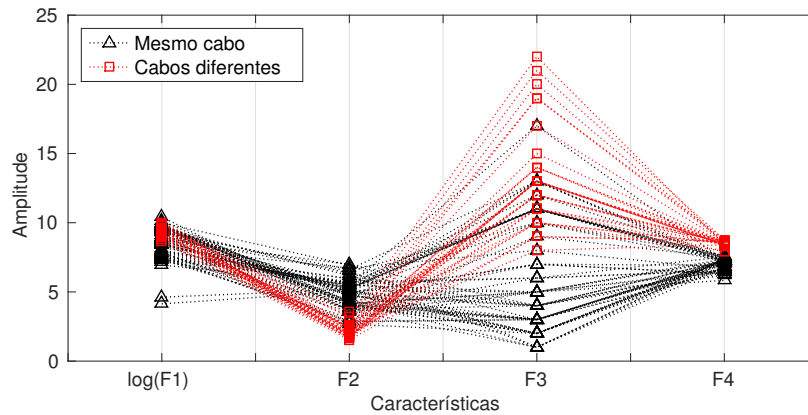


Figura 25 – Vetor característico de 90 amostras S_{11}^{PM} da base de dados considerando um problema de identificação binário de *binders*. Mesmo *binder* e *binders* diferentes são consideradas como apenas uma classe.

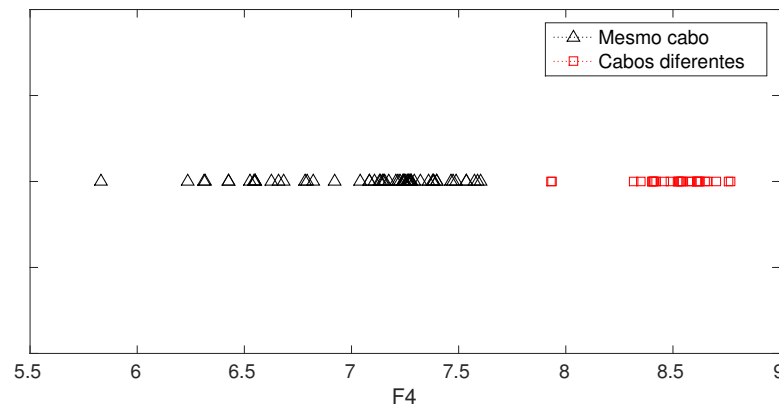


Figura 26 – Visualização da f_4 para 90 amostras S_{11}^{PM} selecionadas randomicamente da base de dados. Uma separação clara das duas classes pode ser vista nesta figura.

algumas características são melhores que outras. Por exemplo, a característica f_2 e f_4 claramente demonstram divisões melhores do que as características f_1 e f_3 . Outra visão sobre a eficiência da característica f_4 , primeiro ponto da TDR (NER), é apresentada na Figura 26, onde é possível visualizar claramente a separação das duas classes. Esta separação demonstra a alta sensibilidade da f_4 para as medições fantasmas.

As próximas duas figuras 27 e 28 discutem a identificação de *binders* como sendo um problema ternário, i.e., considerando todas as três classes do problema original. A primeira figura ilustra como as quatro características se comportam para cada medida S_{11}^{PM} . Percebe-se que neste cenário o reconhecimento de padrão torna-se uma tarefa mais complicada e a separação de classes não é tão clara, onde outrora mostrava-se mais fácil (Figura 25).

Na segunda, Figura 28 (parte superior da figura), destaca apenas a característica f_2 para ser analisada com mais detalhes. A classe mesmo *binder* apresenta um intervalo

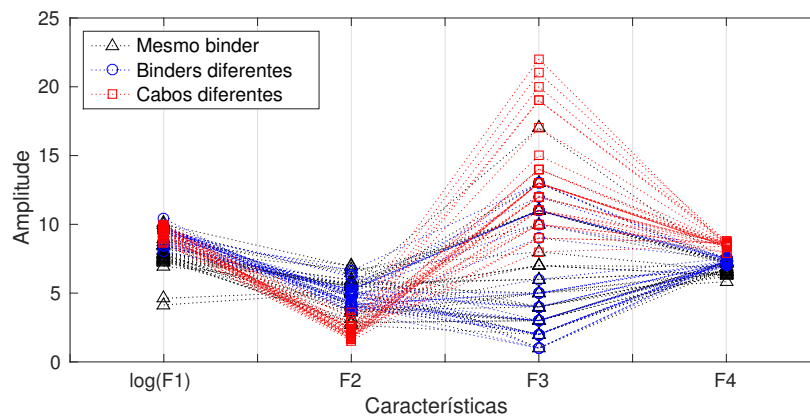


Figura 27 – As mesmas 90 amostras S_{11}^{PM} . Agora considerando a identificação de *binders* como um problema ternário com 30 amostras de cada classe.

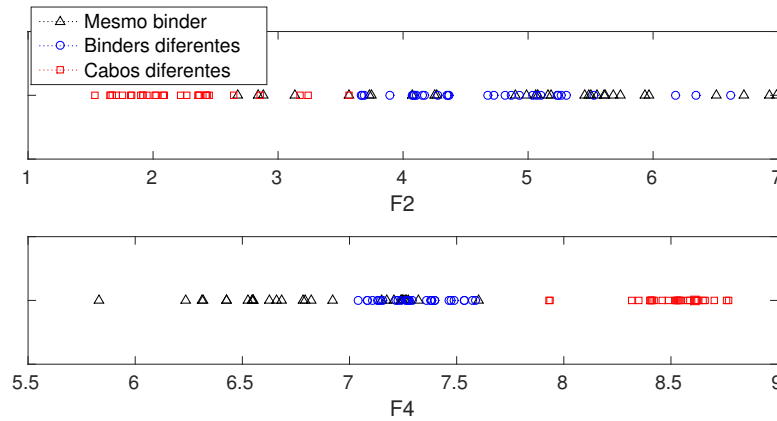


Figura 28 – Vetor característico de 30 medições S_{11}^{PM} para cada classe: (parte superior) Destaca a característica f_2 apenas. (parte inferior) Destaca a característica f_4 apenas.

grande de valores para esta característica (a afirmação também é verdadeira para as características f_1 e f_2), já mostrando desde então que o problema ternário de identificação de *binders* é um tarefa mais complexa para as técnicas de reconhecimento de padrão. A Figura 28 (parte inferior da figura) considera apenas a característica f_2 , revela-se, pelas amostras, que o principal problema recai na diferenciação das classes de mesmo *binder* e *binders* diferentes, onde as amostras da classe cabos diferentes se mantém comportadas com os valores mais a esquerda. No capítulo de resultados, ver-se-á através das acurácias dos modelos a mesma dificuldade de classificar corretamente esses dois padrões.

4.5 Método de Rotulação de *Clusters*

Quando se trabalha com estratégias de aprendizado não supervisionado uma observação importante deve ser feita. Para se calcular a acurácia de um modelo em uma tarefa de classificação, deve-se comparar a classe real de uma amostra com a classe predita pelo modelo. Considere a Equação 4.3 que descreve um cálculo de acurácia para uma base de dados de tamanho m .

$$acc = \frac{1}{m} \sum_{i=1}^m \Omega(\hat{y}_i, y_i) \quad (4.3)$$

onde $\Omega(\cdot, \cdot)$ é uma função binária que compara a classe real y_i da amostra com a classe predita \hat{y}_i , sujeito a

$$\Omega(\hat{y}_i, y_i) = \begin{cases} 1, & \therefore \hat{y}_i = y_i \\ 0, & \therefore \hat{y}_i \neq y_i \end{cases} \quad (4.4)$$

Para se obter a classe \hat{y} predita de todas as amostras, o classificador deve ser capaz de classificá-las mediante critérios aprendidos ao longo de seu treinamento. No algoritmo K -means, por exemplo, cada *cluster* recebe um identificador $j \in [1, 2, \dots, k]$ e

cada amostra é classificada atribuindo-lhe também um identificador, na forma

$$C_i = \arg \min_{j \in [1, 2, \dots, k]} \|f(S_{11}^{PM}) - w_j\|^2. \quad (4.5)$$

Note que C_i é o identificador do centroide mais próximo a amostra $f(S_{11}^{PM})$. Este identificador não remete a uma classe do problema propriamente dito, pois a priori os *clusters* formados pelo modelo não possuem significado lógico, i.e., não há relação direta entre os *clusters* descobertos e as classes do problema. Para isso, é necessário o auxílio de um especialista que reconheça os padrões formados pelas respectivas densidades de amostra da base dados e, conseqüentemente, atribuir-lhes rótulos.

A interpretação de *clusters* não é uma tarefa trivial. Principalmente em problemas de monitoração onde o sistema perpassa por vários estados ao longo do tempo. Portanto, cita-se aqui alguns fatores que dificultam a rotulação: número de padrões a serem reconhecidos pelo algoritmo, número de características utilizadas, sensibilidade de cada característica, nível de conhecimento do especialista sobre o problema, complexidade do problema, tipo de algoritmo utilizado e, principalmente, da combinação das características utilizadas para construir o modelo.

De acordo com a problemática descrita acima e visto que este trabalho remete a identificação – automática – de *binders*, para estar de acordo com os objetivos, deve-se desenvolver um mecanismo para embarcar o conhecimento do especialista aos modelos de clusterização produzidos. Por conseqüente, desenvolveu-se um método para rotular automaticamente os *clusters* obtidos pelas algoritmos de clusterização. Este método utiliza-se primordialmente das informações obtidas através da Seção 4.4.

Uma das informações mais relevante encontrada no estudo das características foi o comportamento das amostras quando observa-se apenas a característica \mathcal{A}_1 , Figura 28 (baixo). Esta característica está diretamente relacionada com o efeito NER que diz respeito ao descasamento de impedância entre o equipamento de medição e o circuito fantasma. A informação chave está na disposição das amostras em relação a esta característica. A figura mostra uma assinatura bem definida para cada classe, mantendo amostras da classe cabos diferentes com valores mais a direita e amostras da classe mesmo *binder* mais a esquerda. Portanto, os *clusters* podem ser imediatamente rotulados se, e somente se, o modelo for gerado apenas por esta característica \mathcal{A}_1 .

Os rótulos dos modelos construídos apenas com a característica \mathcal{A}_1 são automaticamente definidos na ordem decrescente de valores dos centroides, i.e., em relação ao posicionamento dos centroides no espaço característico. O centroide que aglomera em um *cluster* as amostras com maior efeito NER, recebe a rotulação de cabos diferentes e, de forma decrescente no valor do efeito, a rotulação dos outros dois centroides: para efeito NER intermediário rotula-se como *binders* diferentes, e de menor efeito NER como mesmo *binder*. Esse conhecimento é igualmente útil quando um modelo é construído utilizando-se

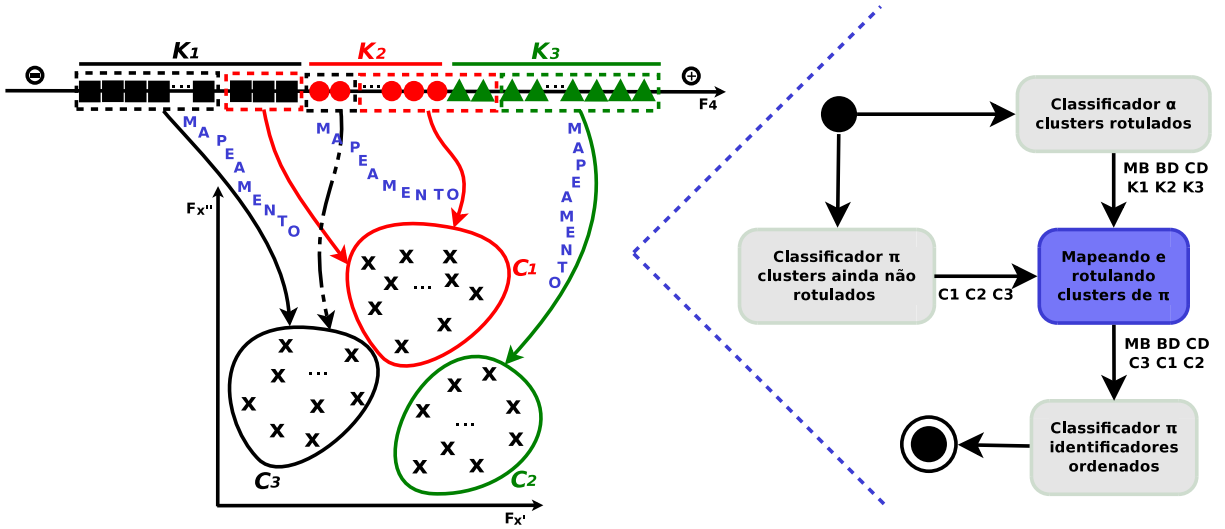


Figura 29 – Representação visual do algoritmo de mapeamento utilizado para rotular *clusters*.

da combinação de outras características, inclusive em casos que a característica \mathcal{A}_1 não está presente no modelo.

O método desenvolvido permite rotular *clusters* construídos a partir de qualquer combinação de características. De maneira minimalista, a Figura 29 ilustra o funcionamento do algoritmo. O exemplo da figura tem a especificidade do problema da identificação de *binder* e, portanto, o número de *cluster* é propositalmente configurado para três. A primeira clusterização constrói o classificador α . Este classificador utiliza a característica \mathcal{A}_1 apenas e, como pode ser visto na figura, os *clusters* $[K_1, K_2, K_3]$ foram encontrados. Sabe-se que os *clusters* são rotulados em ordem decrescente de nível de efeito NER. Logo, a rotulação imediata desta clusterização é da forma descrita na Tabela 1.

Tabela 1 – Rotulação imediata de acordo com a intensidade do efeito NER.

<i>Cluster</i>	Efeito NER	Rótulo
K_3	Alto	cabos diferentes
K_2	Intermediário	<i>binders</i> diferentes
K_1	Baixo	mesmo <i>binder</i>

Em seguida, um outro classificador, denominado de π , é construído a partir da mesma base de dados usada para construir α , exceto que agora esta clusterização é feita utilizando as características $F_{x'}, F_{x''} \in \{\sigma_p^2, \sigma_M^2, n_\Phi, \mathcal{A}_1\}$. E como pode ser visto na figura, os *clusters* $[C_1, C_2, C_3]$ foram encontrados. Como estes *clusters* não foram gerados só, e somente só, através da característica \mathcal{A}_1 , não se pode inferir diretamente os rótulos dos *clusters* do classificador π .

Neste momento, a interação entre o classificador α (já com os *clusters* rotulados) e π (*clusters* ainda não rotulados) ocorre, com o objetivo de rotular os *clusters* do classificador

π . Sabe-se que a mesma base de dados foi usada para construir ambos os classificadores. A diferença está nas características que foram usadas para construí-los. No classificador α , apenas a característica \mathcal{A}_1 foi utilizada, enquanto que no classificador π , as características $F_{x'}$ e $F_{x''}$ foram utilizadas. Portanto, o mapeamento, representado pelas setas na figura, é o procedimento feito para descobrir onde as amostras classificadas pelo classificador α se encontram no espaço característico que contém as mesmas amostras classificadas pelo classificador π .

A seguir, um exemplo de rotulação de um *cluster* do classificador π é apresentado. Para rotular o *cluster* C_1 do classificador π , observa-se que as amostras que estão situadas neste *cluster* são oriundas dos *clusters* K_1 (poucas amostras) e K_2 (grande maioria das amostras) do classificador α (as setas vermelhas indicam a origem das amostras). Neste caso, tem-se dois rótulos candidatos para o *cluster* C_1 do classificador π , que são mesmo *binder* (referente ao rótulo do *cluster* K_1) e *binders* diferentes (referente ao rótulo do *cluster* K_2). A decisão do rótulo escolhido para o *cluster* C_1 é dada através da moda entre os identificadores dos *clusters* K_1 e K_2 . No exemplo, K_2 é eleito como o identificador de C_1 , visto que a grande maioria das amostras é oriunda dele. O fluxograma ao lado apresenta a interação entre classificadores e a Tabela 2 mostra o devido mapeamento entre classificadores, o que corresponde à rotulação dos *clusters* do classificador π .

Tabela 2 – Mapeamento e rotulação através do método proposto.

<i>Cluster</i> de α	Rótulo	<i>Cluster</i> de π
K_3	cabos diferentes	C_2
K_2	<i>binders</i> diferentes	C_1
K_1	mesmo <i>binder</i>	C_3

Desta forma, pode-se rotular *clusters* gerados por quaisquer combinações de características que não envolvam a característica \mathcal{A}_1 e, conseqüentemente, tornar possível o cálculo da acurácia geral do modelo.

O Algoritmo 2 explica sucintamente o procedimento necessário para rotular os *clusters* do classificador π (construído através da combinação de quaisquer características) através do classificador α (construído apenas com a característica \mathcal{A}_1). A Tabela 3 apresenta uma lista de nomes de variáveis com seus respectivos tipos e utilidades, cujo significado é imprescindível para o entendimento do algoritmo.

O método recebe três parâmetros de entrada: o classificador π que terá seus *clusters* rotulados, a base de treinamento \mathbf{T} e o número de *clusters* k . A linha 2 do algoritmo treina um modelo (representado pelo símbolo α) que é construído a partir da base de dados de treinamento \mathbf{T} e o parâmetro de configuração k . Tanto a base de dados, quanto o número de *clusters* são parâmetros externos e devem ser rigorosamente iguais aos usados

Tabela 3 – Lista de variáveis.

Variável	Tipo	Utilidade
\mathbf{T}	matriz $_{p \times n}$: real	base de treinamento com p amostras e n atributos
k	inteiro	número de <i>clusters</i>
π, α	classificador	cada classificador armazena duas estruturas: <i>ids</i> e <i>ctr</i> s
<i>ids</i>	vetor $_p$: inteiro $i \in [1, 2, \dots, k]$	contem a classificação de cada amostra, i.e., o identificador do centroide mais próximo para cada amostra
<i>ctr</i> s	matriz $_{k \times n}$: real	contem o posicionamento de cada centroide
<i>MB_rotulo_α</i>	inteiro	rótulo da classe mesmo <i>binder</i> para o classificador <i>alpha</i>
<i>CD_rotulo_α</i>	inteiro	rótulo da classe cabos diferentes para o classificador <i>alpha</i>
<i>MB_inds</i>	vetor $_p$: binário	índices (do vetor <i>ids</i>) de amostras classificadas como de mesmo <i>binder</i>
<i>CD_inds</i>	vetor $_p$: binário	índices (do vetor <i>ids</i>) de amostras classificadas como de cabos diferentes
<i>CD_rotulo_π</i>	inteiro	rótulo da classe cabos diferentes para o classificador π
<i>MB_rotulo_π</i>	inteiro	rótulo da classe mesmo <i>binder</i> para o classificador π
<i>BD_rotulo_π</i>	inteiro	rótulo da classe <i>binders</i> diferentes para o classificador π

no processo de treinamento do classificador π . Note que apenas a característica \mathcal{A}_1 da base de dados \mathbf{T} é usada.

Nas linhas [3 – 4] são descobertos os índices (ou rótulos) do vetor *ctr*s que se encontram o menor e maior valor, respectivamente. A linha 5 resulta em um vetor binário de mesmo tamanho da base de treinamento, contendo o inteiro 1 em todas as posições em que *ids* é igual ao rótulo pesquisado, e 0 caso contrário. O mesmo procedimento é feito para a linha 6 referente ao outro rótulo pesquisado.

As linhas [7 – 13] realizam o mapeamento de rótulos entre os classificadores. A informação chave é descobrir em que *cluster* do classificador π se encontra cada amostra classificada pelo classificador α . Esta informação é obtida através do mapeamento. As linhas [7 – 8] mapeiam as amostras através da operação $\pi(\cdot)$. Esta operação funciona tal

Algoritmo 2: Rotulando *clusters* com conhecimento indireto da informação contida na característica \mathcal{A}_1 .

Entrada: π, \mathbf{T}, k
Saída: Classificador π com *clusters* rotulados

```

1 início
2    $\alpha := \text{Treinamento}(\mathbf{T}, k);$ 
3    $MB\_rotulo\_alpha := \text{Min\_idx}(\alpha.ctr);$ 
4    $CD\_rotulo\_alpha := \text{Max\_idx}(\alpha.ctr);$ 
5    $MB\_inds := (\alpha.ids = MB\_rotulo\_alpha);$ 
6    $CD\_inds := (\alpha.ids = CD\_rotulo\_alpha);$ 
7    $CD\_rotulo\_pi := \text{Moda}(\pi.ids(CD\_inds));$ 
8    $MB\_rotulo\_pi := \text{Moda}(\pi.ids(MB\_inds));$ 
9   se  $CD\_rotulo\_pi = MB\_rotulo\_pi$  então
10    |  $inds := \pi.ids(MB\_inds);$ 
11    |  $MB\_rotulo\_pi := \text{Moda}(inds(inds \neq CD\_rotulo\_pi));$ 
12  fim
13   $BD\_rotulo\_pi := \text{RotuloFaltante}(CD\_rotulo\_pi, MB\_rotulo\_pi);$ 
14   $\pi := \text{Organiza}(\pi, MB\_rotulo\_pi, BD\_rotulo\_pi, CD\_rotulo\_pi);$ 
15 fim
16 retorna  $\pi$ 

```

qual o exemplo descrito a seguir:

$a := [3, 2.5, 7, 5, 1]$	\mathbf{a} é um vetor
$b := [1, 0, 1, 0, 1]$	\mathbf{b} é um vetor binário
$c := a(b)$	\mathbf{c} é um vetor igual a $[3, 7, 1]$

O resultante desta operação $\pi(\cdot)$ é um vetor de inteiros contendo a classificação de amostras no espaço característico de π . Note que não são todas as amostras que são mapeadas, mas sim apenas aquelas convenientes representadas pelo inteiro 1 no vetor binário usado pelo argumento da operação $\pi(\cdot)$. A moda matemática calcula o rótulo que mais se repete no vetor resultante da operação $\pi(\cdot)$ e, finalmente, obtém-se o rótulo do *cluster* no espaço característico de π .

Existe a possibilidade de um mesmo *cluster* do classificador π ter mais de um rótulo. A linha 9 detecta este problema comparando os rótulos identificados até então. Se a assertiva for verdadeira, um novo mapeamento é realizado para procurar o segundo rótulo que mais se repete no vetor *inds* obtido na linha 10. A operação da linha 11 funciona tal qual o exemplo a seguir:

$a := [3, 2, 1, 2, 2]$	\mathbf{a} é um vetor
$b := 2$	\mathbf{b} é um inteiro
$c := a \neq b$	\mathbf{c} é um vetor binário igual a $[1, 0, 1, 0, 0]$
$d := a(c)$	\mathbf{d} é um vetor igual a $[3, 1]$

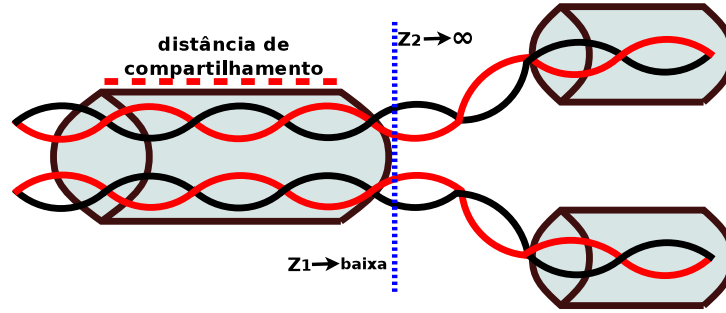


Figura 30 – Dois PTs compartilhando o mesmo cabo e posteriormente se dividindo em cabos diferentes. O ponto em que dois PTs se dividem tem comportamento de circuito aberto. A distância de compartilhamento representa o comprimento que eles compartilham o mesmo cabo.

O resultante desta operação $inds(\cdot)$ é um vetor de inteiros contendo a classificação de amostras no espaço característico de π , excetuando um dos rótulos repetidos. Novamente, a moda matemática é calculada afim de descobrir o (“segundo”) rótulo que mais se repete. Por exclusão, a linha 13 define o rótulo que falta para a classe *binders* diferentes. E, finalmente, na linha 14 rotulam-se os *clusters* do classificador π .

4.6 Identificação do Comprimento de Compartilhamento

A técnica LTI aplicada neste trabalho é usada para determinar a distância de compartilhamento de dois PTs no mesmo cabo. O método proposto é uma versão mais simples do sistema especialista descrito em (47, 48), já que no método para calcular o comprimento de cabos não há necessidade de estimar o diâmetro do condutor. Na Figura 30 tem-se a representação da distância de compartilhamento que é estimada pelo algoritmo proposto. A distância de compartilhamento é calculada através da descoberta do ponto em que os PTs se dividem em cabos diferentes. Neste ponto, há uma forte reflexão do sinal no domínio do tempo, sendo possível coletar o tempo total que o sinal demorou para viajar pelo condutor e ser refletido ao equipamento de medição. Esse tempo serve para estimar o comprimento do cabo através da equação

$$d = 0.67c \times \frac{t}{2}, \quad (4.6)$$

onde o valor $0.67c$ representa uma aproximação da velocidade da corrente elétrica no cobre com velocidade da luz $c = 300.000 \text{ km/s}$ (49) e t é o tempo que o sinal demorou para percorrer todo o comprimento do cabo e ser refletido ao equipamento de medição.

Uma característica importante de dois PTs que formam o canal fantasma é que existe um forte acoplamento entre eles quando se encontram no mesmo *binder*. Isso faz com que o circuito fantasma se comporte como um sinal de par trançado comum, i.e., a impedância característica do canal fantasma (representado por Z_1 na Figura 30) se torna similar ao sinal em modo diferencial de um par trançado e também se assemelha

à impedância do equipamento de medição. Entretanto, quando esses PTs se dividem em *binders* ou cabos diferentes, o acoplamento entre eles diminui e a impedância característica (Z_2) neste ponto é semelhante a um circuito aberto. Uma forte reflexão pode ser notada no domínio do tempo e a distância de compartilhamento (representado pelo símbolo d na Equação 4.6) pode ser calculada. Geralmente, esta situação ocorre quando os PTs se aproximam das dependências do cliente. Portanto, esta estimativa de comprimento também ajuda a distinguir o que pertence as dependências da operadora daquilo que pertence as dependências do cliente.

O Algoritmo 3 apresenta o passo a passo necessário para calcular a distância de compartilhamento. O método recebe três argumentos: a medição em modo fantasma S_{11}^{PM} e dois limiares. A medição S_{11}^{PM} de um circuito fantasma é convertida para o domínio do tempo através da transformada inversa de Fourier (linha 2). O método aplica um procedimento de estimativa para determinar a localização de todos os picos do sinal baseado na variação das técnicas descritas em (47, 48), conhecida como transformada *wavelet* (linha 3). Essa transformada retorna todas as singularidades do sinal. Uma singularidade matemática é o ponto em que uma objeto matemático não é definido, e.g., um ponto de função não diferenciável. Em processamento de sinais, as singularidades marcam pontos importantes de uma TDR, como por exemplo os ecos. A Figura 31 apresenta um exemplo de singularidades detectadas de um sinal em modo fantasma no domínio do tempo. Até esta etapa do algoritmo, ainda não foi descoberto quais bordas são de subida e descida.

Na linha 4, as bordas (singularidades) são separadas em *bSub* (bordas de subida) e *bDes* (bordas de descida) através da verificação de amplitudes de seus vizinhos, onde

$$S_{11}^{PM}(t-1) < S_{11}^{PM}(bordas) < S_{11}^{PM}(t+1) \rightarrow bSub; \text{ e}$$

$$S_{11}^{PM}(t-1) > S_{11}^{PM}(bordas) > S_{11}^{PM}(t+1) \rightarrow bDes.$$

Algoritmo 3: Algoritmo para estimar o comprimento de pares trançados.

Entrada: $S_{11}^{PM}(f)$, lh , lv
Saída: Distância d

- 1 **início**
- 2 $S_{11}^{PM}(t) := \mathbf{Ifft}(S_{11}^{PM}(f));$
- 3 $bordas := \mathbf{Wavelet}(S_{11}^{PM}(t));$
- 4 $[bSub, bDes] := \mathbf{Detectar}(bordas);$
- 5 $[fbSub, fbDes] := \mathbf{Filtrar}(bSub, bDes);$
- 6 $picos := \mathbf{Parear}(fbSub, fbDes, lh, lv);$
- 7 $pico := \mathbf{EncontrarPicoCB}(picos);$
- 8 $d := \mathbf{CalcularDistancia}(pico);$
- 9 **fim**
- 10 **retorna** d

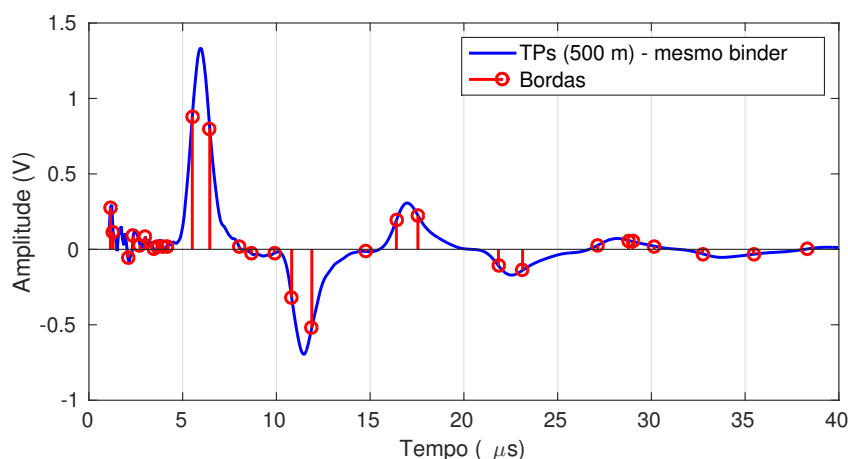


Figura 31 – Exemplo de singularidades encontradas em um sinal em modo fantasma.

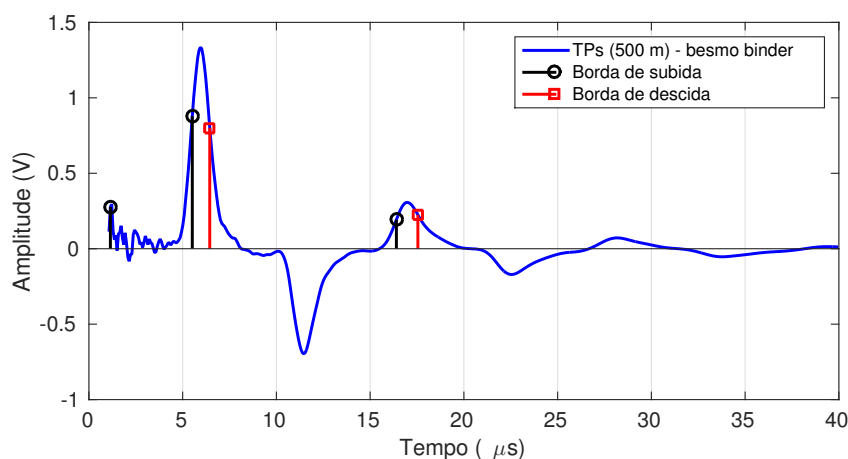


Figura 32 – Exemplo da aplicação do filtro de bordas em um sinal em modo fantasma.

Ainda na mesma figura, percebe-se que algumas bordas são negativas e outras delas são picos espúrios causados por múltiplas reflexões do sinal no domínio do tempo. Para eliminar estes dois tipo de borda, utiliza-se um filtro na linha 5. Eliminar bordas negativas é trivial. Já para eliminar pontos espúrios requer um parâmetro de configuração, i.e., um limiar que regula a amplitude aceitável de um pulso, de tal forma que bordas de amplitude abaixo do limiar são eliminadas dos vetores $bSub$ e $bDes$. A Figura 32 mostra o resultado do filtro aplicado ao sinal fantasma de exemplo. Observe que ainda é possível visualizar bordas que não servem para a estimativa de comprimento, estas bordas são eliminadas naturalmente com a aplicação do próximo passo (linha 6).

O método responsável por parear as bordas de subida e descida, i.e., detectar qual borda de subida está associada a borda de descida, faz uso, além das bordas, de mais duas outras variáveis, que são: lh (limiar horizontal) e lv (limiar vertical). A primeira variável valida o comprimento de um pulso. Se a distância ($distSubDes$) entre a borda de subida e

a borda de descida satisfazer

$$cp - cp \times lh < distSubDes < cp + cp \times lh,$$

onde cp é uma constante configurada de acordo com a largura de pulso de entrada do sinal, então a validação horizontal é satisfeita. Para todas as validações horizontais satisfeitas, uma segunda verificação deve ser feita em relação a amplitude do pulso. O segundo limiar valida a amplitude de um pulso baseado na amplitude da borda de subida. Portanto, o pareamento ocorre quando a borda de descida

$$bSub - bSub \times lv < fbDes < bSub + bSub \times lv.$$

Um exemplo de saída deste método pode ser visto na própria Figura 32, desconsiderando apenas a primeira borda de subida, visto que esta borda não se associará a nenhuma borda de descida próxima a ela. Ao final, restarão apenas duas bordas de subida e duas de decida corretamente associadas.

Os limiares, lh e lv foram automaticamente definidos previamente através de uma pequena base de dados contendo sinais em modo fantasma. Esta base contém o comprimento real dos *binders* onde todas medições foram feitas. Os limiares são definidos através de exaustivos testes de valores para lh e lv . O intervalo testado para os limiares foi de $[0.5, 2.5]$ e a métrica de avaliação para selecionar os melhores limiares é baseada no erro médio calculado entre o valor real do comprimento do *binder* e o estimado. Quanto menor for o erro médio, melhor são os valores de limiares escolhidos. Para a base de dados trabalhar nesta dissertação, os valores dos limiares foram fixados em $lh = 1.3$ e $lv = 1.3$.

Perceba que, pelo exemplo dado, tem-se apenas dois picos restantes representados pelas suas respectivas bordas de subida e decida. O objetivo do método da linha 7 é selecionar apenas o pico referente à mudança brusca de impedância característica do canal fantasma, i.e., quando os PTs que inicialmente estão situados no mesmo *binder* se separam. Este ponto é equivalente a um circuito aberto, tal qual descrito na Figura 30. O método retorna apenas o pico referente ao ponto de distanciamento entre os PTs chamado pelo método de “Pico de Comprimento de *Binder*” (ou simplesmente “PicoCB”). O pico selecionado é aquele que possui o maior valor entre os restantes. A saída deste método é o equivalente mostrado na Figura 33. Finalmente, o comprimento que os PTs compartilham no mesmo *binder* é calculado através da localização deste pico pela Equação 4.6 (linha 8).

Ainda sobre o algoritmo para identificar o comprimento de compartilhamento, é importante destacar também o comportamento diferente do sinal nos dois cenários possíveis de serem avaliados pela técnica que estima comprimentos (Figura 34). Percebe-se que no cenário de PTs no mesmo *binder*, o sinal S_{11}^{PM} apresenta comportamento de maior periodicidade em relação ao cenário de *binders* diferentes. Um fator que contribui fortemente para este comportamento é a impedância característica do canal fantasma.

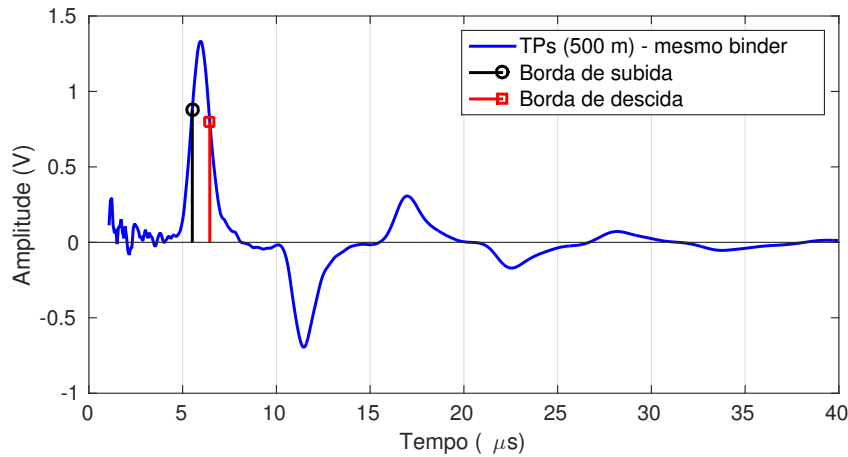


Figura 33 – A resposta no domínio do tempo para dois PTs no mesmo *binder*. O círculo e o quadrado são os pontos de subida e descida do pulso, respectivamente.

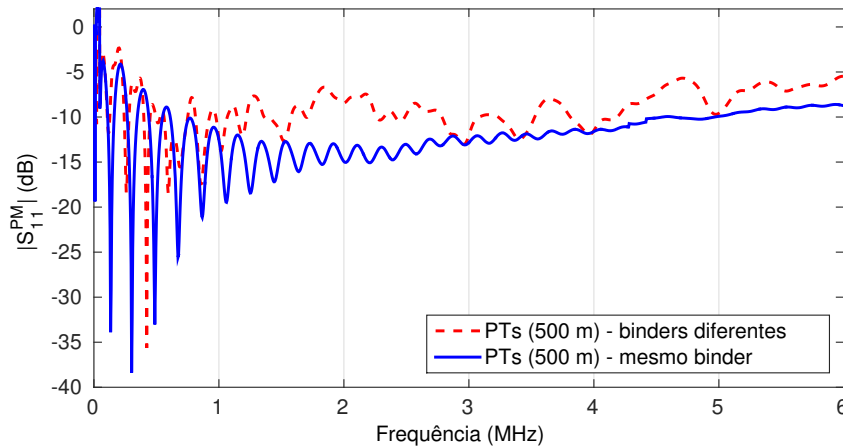


Figura 34 – Duas medições S_{11}^{PM} : PTs no mesmo *binder* (linha cheia) têm alta periodicidade $|S_{11}^{PM}|$ e PTs em *binders* diferentes (linha tracejada) têm baixa periodicidade.

A impedância característica do circuito fantasma formado por PTs no cenário de mesmo *binder*, é baixa. A impedância característica do equipamento de medição também é baixa. O que ocorre é que como ambas impedâncias têm valores próximos, o descasamento de impedância é baixo. Isso faz com que boa parte do sinal fantasma propague ao longo da linha, pois o efeito NER é também baixo. Essa energia inserida na linha contribui para criar fortes ondas estacionárias no domínio da frequência (linha cheia na figura supracitada).

Circuitos fantasmas formados por PTs situados em *binders* diferentes têm impedância característica alta, o que faz com que parte do sinal seja refletido logo na entrada do equipamento de medição e outra parte do sinal se propague ao longo da linha. Isso ocorre devido ao alto descasamento de impedância entre o circuito fantasma e o equipamento de medição, este que possui baixa impedância característica. Neste cenário o sinal perde periodicidade, entretanto ainda mantém periodicidade com fracas ondas estacionárias (linha tracejada na figura supracitada).

Como citado anteriormente e observado pela figura, no cenário de mesmo *binder* o comportamento do sinal fantasma se assemelha ao comportamento de um sinal que foi medido em apenas um PT em modo diferencial, por esta razão a técnica LTI deste trabalho encontra melhores resultados neste cenário. Os resultados sobre a estimação de comprimento são mostrados na Seção 5.5.

5 Resultados

Este capítulo apresenta os resultados obtidos através da aplicação dos algoritmos de aprendizado de máquina à base de dados que contém medições em modo fantasma. Também são apresentados os resultados do método que estima o comprimento de *binders*. Como já discutido previamente na Seção 3.4, a base de dados contém a classificação real de cada amostra. Esta informação é usada apenas em dois momentos: i) para calcular a acurácia do classificador construído pelos algoritmos (de acordo com a Seção 4.5); e ii) para executar o algoritmo SVM que utiliza abordagem supervisionada. Nos algoritmos *K*-means e GMM a base de dados é trabalhada de acordo com as premissas de um aprendizado não supervisionado, portanto a informação da classe que cada amostra pertence não é levada em consideração.

O estudo das combinações de características para cada algoritmo também é apresentado neste capítulo. A conclusão desse estudo é usado como critério para escolher a melhor combinação de características encontradas na base de dados. Um estudo mais específico, considerando apenas a melhor combinação de características, é aplicado à base de dados. Este estudo tem propósito de se aprofundar ainda mais nos resultados obtidos pelos classificadores. Para tal, requer que o melhor classificador construído por cada algoritmo seja aplicado em toda base de dados. Neste capítulo, ainda, uma análise estatística de erro individual para os dois cenários (medições no mesmo *binder* e em *binders* diferentes) é discutida, sobressaltando os principais motivos que levam um cenário a ser mais fácil de estimar o comprimento de *binders* do que o outro cenário.

5.1 Cenário de Medição, Coleta e Transformação dos Dados

O cenário de medição foi escolhido para representar situações em que um par trançado se encontra relacionado com outro par trançado. Neste trabalho, *binder* significa um conjunto de TPs que são trançados entre si. Cabos são constituídos de um conjunto de *binders*. As medições fantasmas são feitas através de dois pares trançados conectados ao equipamento *Network Analyzer* (NA) através de um transformador.

O *setup* de medição é semelhante ao utilizado para realizar medições de parâmetro de espalhamento em pares trançados individuais, com diferença no adicional de um transformador *balun*, que converte o sinal em modo comum para o modo diferencial. A Figura 35 apresenta os elementos participantes do *setup* de medição: 1) o computador executa instruções para que o NA emita um sinal em modo comum; 2) o NA se encarrega de injetar o sinal no meio físico; 3) o transformador *balun* converte o sinal em modo comum oriundo do NA para o modo diferencial; 4) os dois PTs estão conectados em paralelo na

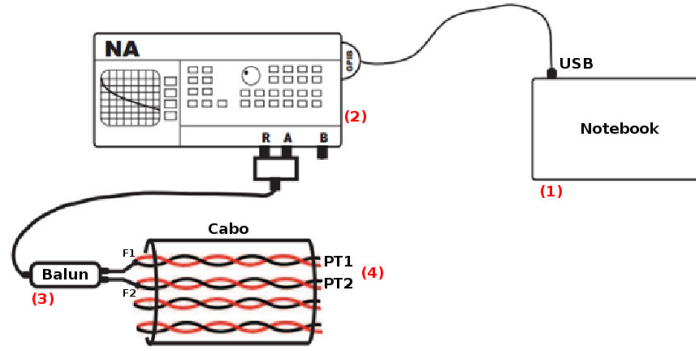


Figura 35 – *Setup* para medição fantasma de uma porta. O *Network Analyzer* é controlado pelo computador e gera sinal em modo comum que é convertido para sinal em modo diferencial conectado em paralelo em ambos os PTs.

saída do *balun* como se fossem dois fios, de modo que a parte positiva do sinal em modo diferencial é injetado no PT1 e a parte negativa do sinal é injetado no PT2. Desta forma, cada PT é visto como um fio, e representados por F1 e F2 na figura. A Equação 5.1 mostra como calcular o parâmetro de espalhamento de uma porta do circuito fantasma.

$$S_{11}^{PM} = \frac{V_{PM}^-}{V_{PM}^+}, \quad (5.1)$$

onde V_{PM}^- e V_{PM}^+ representam o sinal recebido e enviado, respectivamente.

As medições foram realizadas considerando várias combinações de PTs disponíveis no laboratório. Foram coletados 267 sinais fantasmas e armazenados na base de medições **A**. Ao todo, 164 PTs diferentes aglomerados em *binders* foram utilizados. Os *binders* estão organizados em dois modelos de cabos: TEL 313 000 ELQXBE e TEL 481 02 ELAFQBU/120. O primeiro é um cabo não blindado composto por 3 *binders*, cada *binder* contendo 10 PTs. O segundo é um cabo blindado com apenas um *binder* contendo 16 pares.

A transformação da base de medições **A** em **B** ocorre pela extração de característica $f(S_{11}^{PM})$ de cada sinal fantasma coletado (detalhes na Seção 4.2). A base de dados $\mathbf{B} \in \mathbb{R}^{267 \times 4}$, é uma matriz onde cada dimensão armazena um resultado respectivo a extração de características para o sinal. A base de dados contém medições de todos os cenários, a considerar: *intra-binder* (mesmo cabo e no mesmo *binder*), *inter-binders* (mesmo cabo e em *binders* diferentes) e *inter-cabos* (cabos diferentes). Para cada cenário tem-se 89 medições de sinal fantasma, formando um banco de dados balanceado. O comprimento dos PTs variam de 150 – 500 m, típico de aplicações recentes como *G.fast* e *FemtoWoC*.

5.2 Resultados da Máquina de Vetores de Suporte

Dito anteriormente, o algoritmo SVM trabalha em problemas binários de classificação, onde as únicas classes possíveis são compreendidas entre duas: $\{-1, 1\}$. Como trata-se de um algoritmo de aprendizado de máquina supervisionado, cabe o interessado mapear o significado de -1 e 1 para o domínio do problema. Como aqui neste trabalho tem-se três classes (mesmo *binder*, *binders* diferentes e cabos diferentes) em questão, o algoritmo SVM precisará ser executado em duas etapas, i.e., o problema originalmente ternário precisará ser transformado em dois problemas binários. Na primeira etapa (a), considera-se o ambiente como sendo um Problema de Identificação de Cabos (PIC). Já na segunda etapa (b), considera-se como sendo um Problema de Identificação de *Binders* (PIB). Na etapa (a), mapeia-se o PIC conforme:

$$\begin{aligned} \text{mesmo } binder \cup binders \text{ diferentes} &\rightarrow \text{mesmo cabo} \rightarrow -1; \\ \text{cabos diferentes} &\rightarrow 1, \end{aligned}$$

onde $\{\text{mesmo } binder \cup binders \text{ diferentes}\}$ denota a união das amostras de mesmo *binders* e *binders* diferentes consideradas como pertencentes à classe -1 , e $\{\text{cabos diferentes}\}$ como sendo 1 . Para a etapa (b), o PIB é caracterizado desta forma:

$$\begin{aligned} \text{mesmo } binder &\rightarrow -1; \\ binders \text{ diferentes} &\rightarrow 1, \end{aligned}$$

onde as amostras de mesmo *binder* pertencem à classe -1 e *binders* diferentes à classe 1 . Para que seja feita a classificação de uma única amostra da base de teste deve-se seguir o Algoritmo 1 apresentado na Seção 3.4. O mapeamento inverso, i.e., número $\{-1, 1\}$ para uma das classes, garante a decisão da classe designada à amostra. Os resultados do SVM aplicado ao PIC e PIB são apresentados a seguir.

Para analisar a eficiência do algoritmo SVM na identificação de *binders*, primeiramente optou-se por efetuar uma combinação das características extraídas das medições S_{11}^{PM} para os dois problemas (PIC e PIB). Esta tarefa de combinação tem o objetivo de revelar quais características são mais relevantes para explorá-las com mais detalhes na base de dados \mathbf{B} e, posteriormente, definir quais delas podem ser exclusivamente usadas em cada problema. Uma vez que temos quatro características (atributos) para cada medição na base de dados, a Equação 5.2 resulta no número total de combinações possíveis de características (T_c), dado por

$$T_c = \sum_{i=1}^4 \sum_{j=1}^4 C(4, i) \times C(4, j), \quad (5.2)$$

onde i e j são o número de características usadas para a identificação de cabo e *binders*, respectivamente. Ao todo, pode-se obter T_c igual a 225 resultados, considerando todas

as possíveis combinações distintas de características que podem ser extraídas da base de dados.

A Tabela 4 apresenta todos os resultados das combinações de características para o algoritmo SVM para ambos os problemas (PIC e PIB), utilizando validação cruzada com o número de *folds* $k = 10$. Para facilitar o entendimento desta tarefa, optou-se por apresentar nesta tabela apenas o melhor resultado do conjunto de modelos gerados por i e j , com $C(n, i) \times C(n, j)$ resultados por conjunto. Por exemplo, na linha 1 temos $i = 1$ significando que foi usado apenas uma característica no PIC, e $j = 3$ usando três características no PIB, portanto tem-se um conjunto com $C(4, 1) \times C(4, 3) = 16$ modelos, cada um modelo representando uma das 16 diferentes combinações. Entretanto, apenas a combinação que gerou a melhor acurácia média (também conhecida como *sensibilidade média* ou simplesmente *taxa de acerto*), i.e., $CPIC = [4]$ e $CPIB = [1, 3, 4]$ é mostrada na linha 1. Os modelos foram ordenados pela acurácia (última coluna da tabela).

A melhor combinação de características para o PIC foi $[4]$ (\mathcal{A}_1) e para o PIB foi $[1, 3, 4]$ ($\sigma_p^2, n_\phi, \mathcal{A}_1$), alcançando 95,56% de acurácia para a classe mesmo *binder*, 98,89% de acurácia para a classe *binders* diferentes e 100% de acurácia para identificação da classe cabos diferentes. Note que a acurácia média (última coluna) é dada pela média das acurácias totais de cada modelo gerado pela validação cruzada. Portanto, em média, a acurácia total para a melhor escolha de CPIC e CPIB é 98,15%. O intervalo da média de acurácia (diferença entre a melhor e pior combinação) foi de 6,34%. Este valor caracteriza

Tabela 4 – Média de acurácia para cada combinação de características e classe do algoritmo SVM. $CPIC$ e $CPIB$ significam, respectivamente, Características para o PIC e *Binder* no processo. MB , BD e CD significam a taxa de acerto por classe: Mesmo *Binder*, *Binders* Diferentes e Cabos Diferentes.

#	CPIC	CPIB	MB (%)	BD (%)	CD (%)	Média (%)
1	[4]	[1,3,4]	95,56	98,89	100	98,15
2	[1,2,4]	[1,3,4]	95,56	98,89	98,89	97,78
3	[4]	[1,4]	96,53	96,67	100	97,75
4	[1,4]	[1,3,4]	95,56	98,89	97,78	97,41
5	[1,2,4]	[1,4]	96,53	96,67	98,89	97,38
6	[1,4]	[1,4]	96,53	96,67	97,78	97,01
7	[1,2,3,4]	[1,3,4]	95,56	96,67	97,78	96,67
8	[4]	[1,2,3,4]	89,86	100	100	96,61
9	[1,2,3,4]	[1,4]	96,53	94,44	97,78	96,27
10	[1,2,4]	[1,2,3,4]	89,86	100	98,89	96,24
11	[1,4]	[1,2,3,4]	89,86	100	97,78	95,87
12	[1,2,3,4]	[1,2,3,4]	89,86	97,78	97,78	95,13
13	[4]	[4]	84,31	95,56	100	93,29
14	[1,2,4]	[4]	84,31	95,56	98,89	92,92
15	[1,4]	[4]	84,31	95,56	97,78	92,55
16	[1,2,3,4]	[4]	84,31	93,33	97,78	91,81

uma consistência nos resultados do algoritmo SVM e traz garantia de se ter modelos com acurácia total acima de 90%.

Desta Tabela 4 é possível extrair informações sobre a frequência absoluta de cada característica, incluindo combinações de características aparecendo juntas de dois a dois, três a três e quatro a quatro. Esta análise tem como objetivo identificar a repetibilidade da combinação de características da técnica SVM. A frequência absoluta é expressada por valores inteiros e porcentagens de ocorrência para cada combinação de características nas Tabelas 5, 6 e 7 (ordenadas pela última coluna). A quarta coluna representa a frequência absoluta de cada combinação de características que aparece na segunda coluna como CPIC e na terceira coluna como CPIB. A última coluna representa a frequência absoluta em porcentagem. Note que a fórmula da frequência absoluta remete ao número máximo de ocorrência para cada combinação de características, i.e., um a um pode aparecer até 32 vezes, dois a dois até 24 vezes, três a três até 16 vezes e quatro a quatro até 8 vezes, e.g., na Tabela 6 tem-se a combinação $F_{2,4}$ que aparece 8 vezes como CPIC e 4 vezes como CPIB, então $f_{2,4}$ é igual a 12. E a possibilidade máxima de ocorrência de combinação de características aparecendo de dois em dois é 24, portanto $\frac{f_{i,j}}{24} = \frac{12}{24} = 50$.

Tabela 5 – Frequência absoluta considerando o aparecimento de característica um a um apenas.

F_i	CPIC	CPIB	f_i	$\frac{f_i}{32}$ (%)
[4]	16	16	32	100
[1]	12	12	24	75
[2]	8	4	12	37,5
[3]	4	8	12	37,5

Tabela 6 – Frequência absoluta considerando o aparecimento de característica dois a dois.

$F_{i,j}$	CPIC	CPIB	$f_{i,j}$	$\frac{f_{i,j}}{24}$ (%)
[1,4]	12	12	24	100
[1,2]	8	4	12	50
[1,3]	4	8	12	50
[2,4]	8	4	12	50
[3,4]	4	8	12	50
[2,3]	4	4	8	33,3

Uma observação importante é a presença da característica [4] (\mathcal{A}_1) em todos os melhores resultados apresentados na Tabela 4 e destacado com 100% de frequência absoluta na Tabela 5. Isto ratifica a importância desta característica em separar as amostras no espaço de características, como já foi discutido na Seção 4.4. Já na Tabela 6, as características [1,4] ($\sigma_p^2, \mathcal{A}_1$) têm frequência absoluta elevada entre os resultados, indicando novamente a importância da característica [4] (\mathcal{A}_1) nos resultados. Adicionalmente, na

Tabela 7 – Frequência absoluta considerando o aparecimento de característica três a três.

$F_{i,j,k}$	CPIC	CPIB	$f_{i,j,k}$	$\frac{f_{i,j,k}}{16}$ (%)
[1,2,4]	8	4	12	75
[1,3,4]	4	8	12	75
[1,2,3]	4	4	8	50
[2,3,4]	4	4	8	50

Tabela 7, as características [1,2,4] ($\sigma_p^2, \sigma_M^2, \mathcal{A}_1$) reforçam o aparecimento das duas primeiras características com o surgimento de mais uma (σ_M^2). Outra informação importante, é a presença das características [1,3,4] ($\sigma_p^2, n_\phi, \mathcal{A}_1$) também aparecendo juntas em 75% das vezes (empatando com a primeira combinação da Tabela 7). É esta última combinação que aparece como vencedora em termos de acurácia na Tabela 4. Portanto, deve-se ter uma atenção especial para esta melhor combinação de características. Nenhuma tabela é apresentada para descrever a frequência absoluta para as quatro características aparecendo juntas devido ao óbvio 100% de ocorrência.¹

A Tabela 4 possibilita a descoberta da melhor combinação de características. O objetivo agora é explorar esta melhor combinação na mesma base de dados através de alguns testes. Três testes são realizados na base de dados utilizando apenas a melhor combinação de características gerada pelo algoritmo SVM. Os testes são descritos a seguir:

1. Executa-se o algoritmo SVM com a validação cruzada para *folds* $k = 10$. Cada execução da validação cruzada gera um classificador. Calcula-se a **acurácia total** para cada classificador e a **acurácia média** entre todos os classificadores (mais detalhes sobre o cálculo dessas duas acurácias a seguir).
2. Armazena-se o melhor classificador encontrado pela validação cruzada e destaca-se sua **acurácia total** em relação a base de teste. O resultado deste teste pode ser obtido paralelamente à execução do teste anterior.
3. Executa-se o melhor classificador em toda a base de dados para verificar sua **acurácia geral**.

As diferentes acurácias trabalhadas aqui têm por finalidade avaliar um modelo por diferentes pontos de vista. A acurácia de um modelo que reconhece os três padrões (mesmo *binder*, *binders* diferentes e cabos diferentes) é avaliada separadamente através da sensibilidade que o modelo tem de classificar amostras positivas dentre os padrões, por isso os resultados nas tabelas são apresentados em termos de sensibilidade MB (%), BD

¹ Assume-se a fórmula de porcentagem como sendo $\frac{f_{1,2,3,4}}{8}$, onde **CPIC** e **CPIB** devem ter cada uma 4 ocorrências necessariamente. Isto pode ser afirmado devido a combinação de quatro características ter apenas uma forma possível de aparecerem juntas.

(%) e CD (%). Haja vista que a validação cruzada gera 10 classificadores (teste 1), tem-se as sensibilidades expressadas através da média, chamada aqui de **acurácia média**. Já a **acurácia total** é taxa de acerto de um classificador, em outras palavras, é a relação de verdadeiros positivos pela número de amostras na base de teste. Já a **acurácia geral** é a acurácia total aplicada a base de dados completa.

A Tabela 8 resume as acurácias encontradas nos três testes. As colunas 2, 3 e 4 são as acurácias médias. A última coluna apresenta a acurácia média do teste 1 (linha 2) e a acurácia total do teste 2 (linha 3). Já no teste 3 (linha 4), tem-se a acurácia geral do melhor classificador aplicado a base toda. Como pode ser visto, o algoritmo SVM produz resultados notáveis com acurácia geral de 98,88%, acertando todas as amostras de teste para a classe cabos diferentes 100% e acima de 97% para as outras duas classes.

Tabela 8 – Acurácia do algoritmo SVM para cada classe usando a característica $[\mathcal{A}_1]$ para o PIC e $[\sigma_p^2, n_\phi, \mathcal{A}_1]$ para o PIB.

Tipo	MB (%)	BD (%)	CD (%)	Média (%)
Média dos classificadores	97,78	97,78	100	98,52
Melhor classificador	100	100	100	100
Melhor classificador em toda a base	97,75	98,88	100	98,88

A Figura 36 mostra o melhor classificador SVM para a identificação de binders, onde revela-se o hiperplano que maximiza as margens entre as duas classes. Foi usada um função kernel polinomial com ordem igual ao número de características mais um. A figura mostra um formato de hiperplano 3D separando medições de mesmo *binder* de medições de *binders* diferentes. Esse hiperplano representa a separação binária de amostras em um espaço característico de muitas dimensões. Note que para a classificação de *binders*, uma camada ligeiramente deformada foi suficiente para separar as duas classes. Note também que optou-se por não mostrar o hiperplano para a identificação de cabos visto que este hiperplano certamente seria representado por apenas um traço (ou ponto), uma vez que a melhor combinação de características resultou em apenas uma (\mathcal{A}_1) para o PIC.

5.3 Resultados do *K*-means

Quando se usa um algoritmo de aprendizado de máquina não supervisionado que realiza a tarefa de clusterização de dados, um dos principais problemas a se enfrentar é definir o número de *clusters* que devem ser modelados pelo algoritmo. Este número deve ser inferido pelo usuário do algoritmo. Geralmente não é uma tarefa trivial visto que em muitos problemas do cotidiano o usuário não sabe por quantos estados um determinado sistema passou e, portanto, torna-se uma tarefa difícil especular quantos padrões podem existir na base de dados referente ao sistema. Há na literatura muitos algoritmos de clusterização que fazem esta descoberta automaticamente. Há também diversos critérios que auxiliam

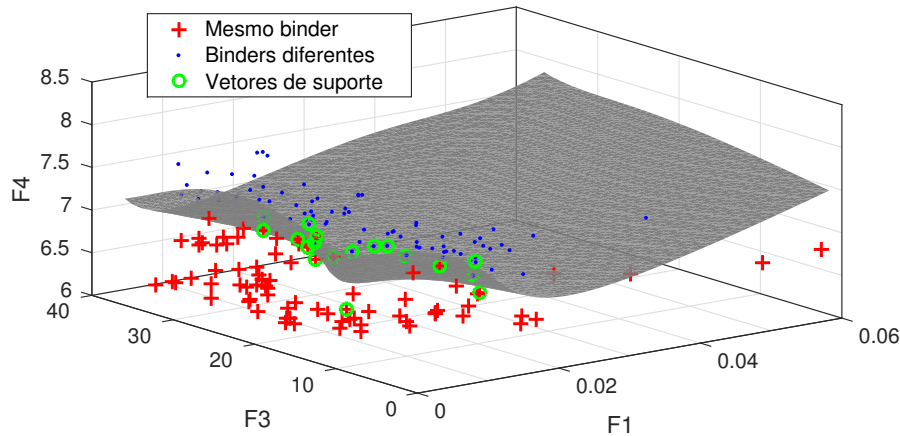


Figura 36 – Hiperplano para identificação de *binders* e vetores de suporte na base de treino.

na descoberta do número ótimo de *clusters*, e.g., *Bayesian Information Criterion* (BIC), *Akaike Information Criterion* (AIC), *Calinski-Harabasz Criterion* (CHC), entre outros. Entretanto, aqui neste trabalho não é necessário utilizar nenhum dos critérios para a escolha do número de *clusters*, pois já se sabe de antemão que o número de *clusters* é escolhido de acordo com o número de classes que se espera separar as amostras da base de treinamento. Neste contexto de identificação de *binders*, o parâmetro k foi configurado para 3, representando exatamente o número de classes: mesmo *binder*, *binders* diferentes e cabos diferentes.

Seguindo os mesmo passos do algoritmo SVM, os resultados do K -means são apresentados a seguir. Primeiramente são apresentados os resultados da combinação de características e em seguida os resultados da melhor combinação de características aplicada à base de dados. É importante ressaltar que este algoritmo não necessita ser executado separadamente, em dois passos, para classificar cabo e *binder*, como ocorre com o SVM. Em apenas um passo, K -means é capaz de construir um classificador que acomode todas as amostras em seus *clusters*. Por este motivo, naturalmente há uma redução no número de combinações possíveis de características, com T_c igual a 15.

A Tabela 9 mostra todas as possíveis combinações de características usando o algoritmo K -means. Alternativamente, apresenta-se também uma outra Tabela 10 que representa um filtro na Tabela 9 contendo apenas as melhores combinações de cada conjunto $C(n, i)$ (o equivalente a Tabela 4 para o SVM). Novamente, estes resultados foram gerados fixando o número de *folds* $k = 10$. A combinação vencedora com a melhor média de acurácia foi com as características [1,4] ($\sigma_p^2, \mathcal{A}_1$), alcançando aproximadamente 92% de acurácia média. Esse classificador teve uma clara dificuldade em identificar medições S_{11}^{PM} oriundas do mesmo *binder*, o que já era esperado devido a análise de características mostrada na Seção 4.4. Entretanto, apesar da porcentagem de erro de aproximadamente 8%, sabe-se

pela análise de características feitas que as amostras de classe mesmo *binder* e *binders* diferentes estão bem próximas uma das outras (visto que se sabe os rótulos das amostras a priori) e, portanto, essa porcentagem de erro é resultante de amostras classificadas falso positivamente para a classe *binders* diferentes. Ainda assim, implica dizer que são amostras que estão no mesmo cabo. Sabe-se disso devido ao resultado da clara separação que a característica [4] (\mathcal{A}_1) proporciona ao identificar medições oriundas do mesmo cabo e cabos diferentes (vide Figura 26).

Tabela 9 – Média de acurácia para cada combinação e classe pelo algoritmo K -means.

#	Características	MB (%)	BD (%)	CD (%)	Média (%)
1	[1,4]	75,42	100	100	91,78
2	[4]	75,42	100	100	91,78
3	[2,4]	59,72	81,94	98,89	80,13
4	[1,2,4]	58,61	80,83	98,89	79,37
5	[1,2]	55,14	79,72	97,78	77,51
6	[2]	54,03	80,97	97,78	77,51
7	[1,2,3]	40,42	29,17	58,33	42,69
8	[3,4]	40,42	31,53	53,89	41,92
9	[3]	39,31	25,83	60,56	41,88
10	[2,3,4]	39,31	28,06	57,22	41,58
11	[2,3]	39,31	31,39	52,78	41,14
12	[1,2,3,4]	39,31	27,08	55	40,39
13	[1,3,4]	40,42	23,75	56,25	40,01
14	[1,3]	40,42	21,25	58,47	40
15	[1]	13,61	1,11	66,25	26,99

Tabela 10 – Média de acurácia das melhores combinações de cada conjunto $C(n, i)$ para algoritmo K -means.

#	Características	MB (%)	BD (%)	CD (%)	Média (%)
1	[1,4]	75,42	100	100	91,78
2	[4]	75,42	100	100	91,78
3	[1,2,4]	58,61	80,83	98,89	79,37
4	[1,2,3,4]	39,31	27,08	55	40,39

Novamente, percebe-se que a característica [4] (\mathcal{A}_1) apenas, aparece em todos as melhores combinações de características do conjunto $C(n, i)$ na Tabela 10. Isto ratifica sua importância e mostra que o efeito NER é um importante indicador no processo de identificação de *binder*. Outra informação importante é que houve um empate de acurácias entre o primeiro e o segundo lugar. Portanto, para a próxima fase de inspeção da melhor combinação de características, optou-se por analisar as duas combinações [1,4] ($\sigma_p^2, \mathcal{A}_1$) e [4] (\mathcal{A}_1).

Aplicando os mesmos três testes descritos nos resultados do SVM em relação às duas combinações vencedoras, K -means produziu os resultados mostrados na Tabela 11

e 12. A tabela confirma que classificar corretamente uma medição S_{11}^{PM} de dois PTs que pertencem ao mesmo cabo é mais difícil do que classificar uma medição S_{11}^{PM} de dois PTs que pertencem a cabos diferentes. Isto pode ser concluído a partir da dificuldade que os modelos produzidos pela validação cruzada encontraram ao classificar amostras de medições S_{11}^{PM} no mesmo *binder* com 75,14% de acurácia média para a combinação [1,4] ($\sigma_p^2, \mathcal{A}_1$) e 75,42 para a combinação [4] (\mathcal{A}_1). Quando aplica-se o melhor classificador em toda a base de dados, *K*-means alcançou quase 92% de acurácia geral com uma classificação perfeita para as classes diferentes cabos e *binders* diferentes, em ambas combinações vencedoras.

Tabela 11 – Média de acurácia para cada classe usando as características $[\sigma_p^2, \mathcal{A}_1]$ para a identificação de cabo e *binder*.

Tipo	MB (%)	BD (%)	CD (%)	Média (%)
Média dos classificadores	75,14	100	98,89	91,39
Melhor classificador	100	100	100	100
Melhor classificador em toda a base	75,28	100	100	91,76

Tabela 12 – Média de acurácia para cada classe usando a característica $[\mathcal{A}_1]$ para a identificação de cabo e *binder*.

Tipo	MB (%)	BD (%)	CD (%)	Média (%)
Média dos classificadores	75,42	100	98,89	91,42
Melhor classificador	100	100	100	100
Melhor classificador em toda a base	75,28	100	100	91,76

Os resultados de acurácia obtidos pelo algoritmo *K*-means são possíveis devido ao método de rotulação de *clusters* (descrito na Seção 4.5) pelo Algoritmo 2. Considerando o empate de duas combinações de características (Tabela 9), demonstra-se através das Figuras 37 e 38 os resultados de classificação de ambos classificadores. O resultado da rotulação dos *clusters* também é visto nas figuras. Pode-se observar que há uma região de interseção entre as amostras de mesmo *binder* e *binders* diferentes, o que faz com que os classificadores se confundam entre essas duas classes, enquanto que as amostras de cabos diferentes estão mais comportadas na região de cima da Figura 37 e a direita da Figura 38 com valores maiores de efeito NER (\mathcal{A}_1).

5.4 Resultados do Modelo de Misturas de Gaussianas

Assim como o algoritmo *K*-means, o algoritmo GMM também não precisa ser executado separadamente para classificar cabo e *binder*. Além disso, os rótulos de cada componente (Gaussiana, ou ainda *cluster*) é rotulado da mesma forma que os *clusters* do algoritmo *K*-means, de acordo com o algoritmo descrito na Seção 4.5. Os resultados apresentados a seguir são oriundos dos mesmos testes realizados com os algoritmos SVM e

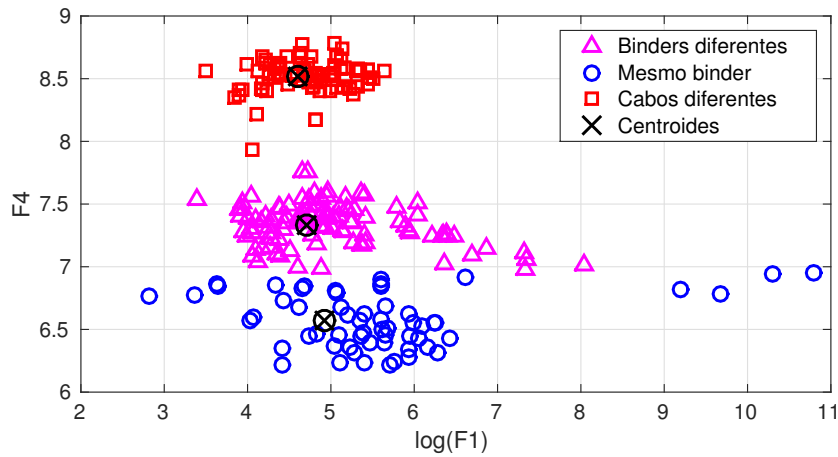


Figura 37 – Resultado da clusterização do algoritmo K -means para o melhor classificador de combinação $[1,4]$ ($\sigma_p^2, \mathcal{A}_1$) aplicado à base de treinamento. Os *clusters* foram devidamente rotulados através da técnica de rotulação descrita neste trabalho.

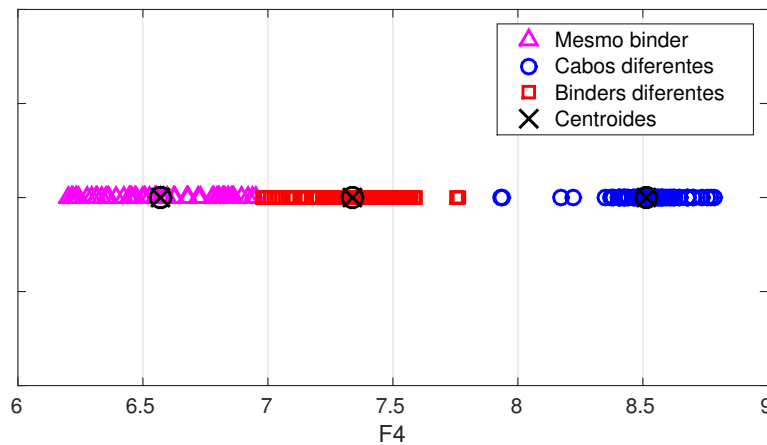


Figura 38 – Resultado da clusterização do algoritmo K -means para o melhor classificador de combinação $[4]$ (\mathcal{A}_1) aplicado à base de treinamento. Os *clusters* foram devidamente rotulados através da técnica de rotulação descrita neste trabalho.

K -means, i.e., primeiro são apresentados os resultados das combinações de características e segundo são apresentados os resultados da melhor combinação aplicados à base de dados.

A Tabela 13 apresenta todos os possíveis resultados de combinação de características. A combinação vencedora utilizou apenas a característica $[4]$ (\mathcal{A}_1), alcançando 92,44% de acurácia média. Note que esta combinação vencedora é a mesma encontrada com o algoritmo K -means (empatado com a primeira posição). Note também que esta combinação vencedora do algoritmo GMM melhorou a acurácia média das medições S_{11}^{PM} de mesmo *binder* em comparação com a apresentada pelo K -means. A Tabela 14 apresenta a melhor combinação de características por conjunto $C(n, i)$.

Seguindo os mesmos três testes aplicados à base pelos algoritmos SVM e K -means, a Tabela 15 apresenta os resultados referentes ao algoritmo GMM. Novamente, a tabela

Tabela 13 – Média de acurácia para cada combinação e classe pelo algoritmo GMM.

#	Característica	MB (%)	BD (%)	CD (%)	Média (%)
1	[4]	82,78	96,67	97,78	92,44
2	[2,4]	68,61	79,31	98,89	82,32
3	[2]	66,39	57,08	96,67	73,38
4	[1,2,4]	76,53	19,72	98,89	64,83
5	[2,3]	38,61	42,64	81,67	54,25
6	[1,2]	74,31	12,36	68,61	51,71
7	[1,4]	35	2,22	100	45,7
8	[2,3,4]	45,69	20,14	67,22	44,53
9	[1,3]	28,89	7,78	96,53	44,53
10	[1,2,3,4]	26,81	37,08	66,11	43,41
11	[1,2,3]	26,81	32,5	68,33	42,65
12	[3,4]	39,03	31,39	55	41,91
13	[1,3,4]	21,25	40,42	62,78	41,54
14	[3]	25,69	35,83	62,78	41,51
15	[1]	7,78	7,78	85,56	33,7

Tabela 14 – Média de acurácia das melhores combinações de cada conjunto $C(n, i)$ para algoritmo GMM.

#	Característica	MB (%)	BD (%)	CD (%)	Média (%)
1	[4]	82,78	96,67	97,78	92,44
2	[2,4]	68,61	79,31	98,89	82,32
3	[1,2,4]	76,53	19,72	98,89	64,83
4	[1,2,3,4]	26,81	37,08	66,11	43,41

Tabela 15 – Acurácia do GMM para cada classe usando a característica \mathcal{A}_1 na identificação de *binder*.

Tipo	MB (%)	BD (%)	CD (%)	Média (%)
Média dos classificadores	83,06	97,78	97,78	92,86
Melhor classificador	100	100	100	100
Melhor classificador em toda a base	84,27	94,38	97,75	92,13

ratifica que classificar corretamente as medições S_{11}^{PM} de mesmo *binder*, que significa dois PTs no mesmo cabo, é mais difícil do que classificar medições S_{11}^{PM} de cabos diferentes. Isto pode ser verificado através da acurácia média individual dos classificadores produzidos pela validação cruzada, da classe mesmo *binder* com 83,06% e da classe cabos diferentes de 97,78%. Quando o melhor classificador é aplicado a base toda, GMM resulta em 92,13% de acurácia geral.

O algoritmo GMM é conhecido como uma técnica de clusterização *soft* ou *fuzzy*, i.e., uma amostra recebe uma probabilidade (posterior) de pertencer a um determinado *cluster*. Diferentemente do algoritmo *K*-means conhecido também como uma técnica de

clusterização *hard*, i.e., uma amostra pertence ou não a uma determinado *cluster*. Esta característica faz com que o algoritmo GMM, apesar de não ter obtido desempenho significativamente melhor em relação ao algoritmo *K*-means, agregar mais informação a sua classificação através da Probabilidade Posterior (PP).

A Figura 39 ilustra as PPs em relação ao *cluster* que remete as amostras de *binders* diferentes, considerando apenas a base de treinamento. Este gráfico torna possível perceber a separação da amostras da classe *binders* diferentes, que apresenta valores altos de PPs, em relação as outras duas classes com valores baixos de PPs. As amostras cujos valores de PPs são baixos significam que têm probabilidade baixa de pertencer a classe *binders* diferentes, pois é em relação a esta classe que essa figura remete. Esta informação garante mais precisão ao operador pois agora, além da classificação de uma amostra, tem-se a probabilidade desta classificação estar correta.

A Figura 40 mostra os valores de PPs de cada amostra em relação a cada componente. As amostras foram ordenada pela PP da classe cabos diferentes. Um bom classificador GMM é aquele que cada amostra da base de dados tenha um alto valor de PP para uma determinada classe, i.e., aquela classe que a amostra realmente pertence, e baixos valores de PPs para as outras classes, i.e., aquelas classes que a amostra não se assemelha ou não pertence. Nesta figura, há apenas três intersecções de linhas entre os três componentes. Isto informa também que poucas amostras ficaram com os valores de PPs comprometidos, i.e., amostras que apresentam PPs semelhantes para cada classe, aumentando a incerteza quanto a classe correta.

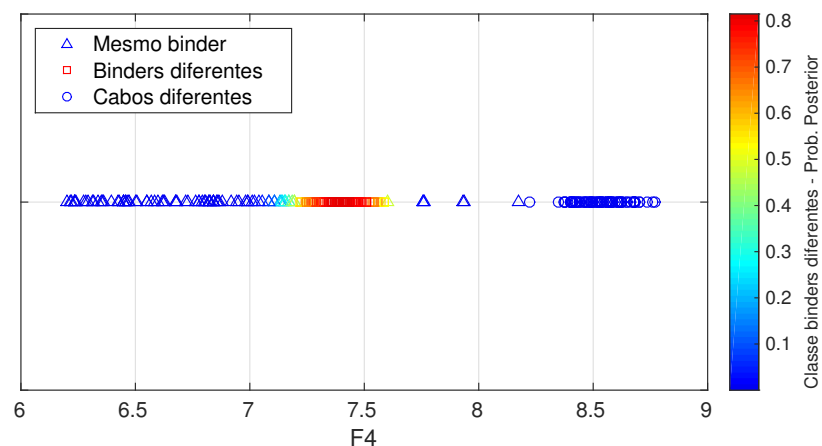


Figura 39 – Resultados do GMM para o melhor classificador usando apenas a característica \mathcal{A}_1 na base de treinamento.

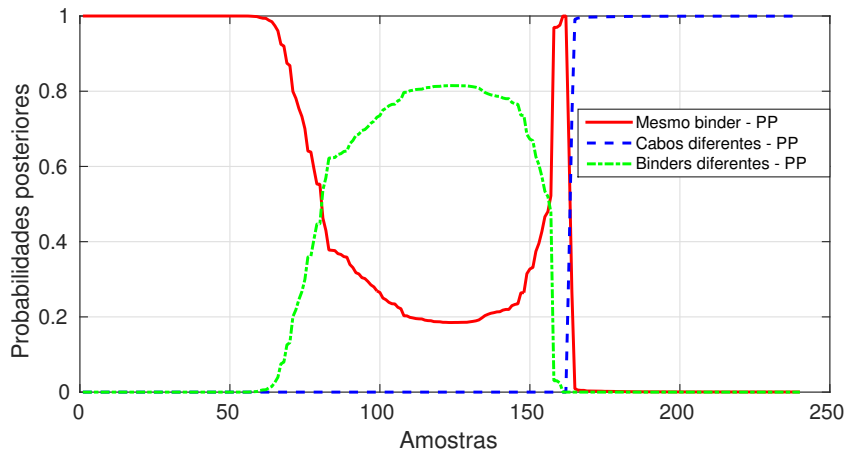


Figura 40 – Probabilidade posterior de cada amostra da base de treinamento. Poucas interseções de curvas indicam um classificador coerente para classificação de dados.

5.5 Resultados da Estimação de Comprimento

Nesta seção será apresentado os resultados da estimação de comprimento de PTs situados no mesmo cabo. A identificação do comprimento de compartilhamento entre dois PTs é estimada pela reflectometria no domínio do tempo de um circuito fantasma através da aplicação da transformada inversa de Fourier em uma medição em modo fantasma no domínio da frequência $S_{11}^{PM}(f)$. Para se analisar estatisticamente os resultados, usou-se a mesma base de dados utilizada no problema de identificação de *binders*. Obviamente, os resultados que veremos a seguir são estimativas de comprimento de medições S_{11}^{PM} cujo os PTs estão no mesmo cabo, que é formado pela medições de mesmo *binder* e *binders* diferentes.

As medições foram coletadas utilizando dois tamanhos distintos de diâmetros de condutor 0.4 mm e 0.5 mm, resolução de frequência de 4,3125 kHz, e intervalo de frequência de até 6 MHz. A Tabela 16 sumariza o banco de dados utilizado composto por 164 medições S_{11}^{PM} , com 80 medições fantasma de mesmo *binder* e 84 de *binders* diferentes. Aproximadamente 61% das linhas têm 500 m de comprimento e em torno de 39% de linhas com 100 m, 150 m, 200 m e 250 m. Estes últimos comprimentos servem para avaliar a técnica de estimativa de comprimento com relação a sensibilidade para distâncias menores. Além disso, sabe-se que tecnologias DSL modernas são susceptíveis a distâncias grandes, diminuindo drasticamente a taxa de transmissão agregada entre origem e destino, por isso a importância de se utilizar comprimentos menores.

A Tabela 17 mostra o resultado da estimativa de comprimento de todas as 164 medições fantasmas. O erro da estimativa de comprimento é calculado através da Equação 5.3. O *erro* é uma métrica que define o quão distante do valor real a estimação se encontra,

Tabela 16 – Visão geral da base de dados para estimação de comprimento.

Tipo	100 m	150 m	200 m	250 m	500 m	Total
Mesmo <i>binder</i>	-	-	6	5	69	80
<i>Binders</i> diferentes	15	24	4	10	31	84

para mais ou para menos.

$$erro (\%) = \pm \frac{|Real - Estimado|}{Real} \times 100 \quad (5.3)$$

A partir da Tabela 17, admite-se um conjunto de informações extraídas com intuito de compilar informações importantes da estimação de comprimento. A Tabela 18, a seguir, apresenta a distribuição de erros entre classes e também o erro total extraído da base de dados. Em linhas gerais, medições S_{11}^{PM} de PTs em *binders* diferentes apresentam um erro médio de 2,86% maior que PTs no mesmo *binder*. O erro médio total de 11,3% está associado a todas as medições fantasmas, independente do comprimento do cabo.

Uma outra visualização da informação pode ser observada na Tabela 19. Esta tabela apresenta os erros por intervalos, e.g., no intervalo em que os erros de estimação ficaram entre]20, 30]%, tem-se 3,66% de amostras, 1,25% das amostras da classe mesmo *binder* e 5,95% das amostras da classe *binders* diferentes. Uma importante informação extraída desta tabela é que para quase 70% das medições de mesmo *binder* o erro foi menor que 10%. Outros 30% restantes estão em intervalos acima de 10%. A resposta da reflectometria no domínio do tempo para medições fantasmas é bem parecida com medições em modo diferencial de pares individuais, o que favorece técnicas LTI com apenas 18,75% de linhas no intervalo]10, 20]%. Por outro lado, o menor acoplamento entre PTs de diferentes *binders* diminui a qualidade da estimativa de comprimento com apenas 30,95% de linhas em *binders* diferentes com erro abaixo de 10%, e quase 70% das linhas situam-se entre o intervalo de]10, 30]%. Não se obteve nenhuma estimativa de comprimento com erro maior que 30% para *binders* diferentes.

A Tabela 20 tem o objetivo de mostrar um erro médio proporcional (equivalente ao “peso” que cada comprimento carrega nas médias dos erros), separado por comprimento de cabo. Este erro é calculado através da equação seguinte

$$emp_j = \frac{em_j}{\sum_{i=1}^t em_i} \quad (5.4)$$

onde emp_j é o erro médio proporcional de comprimento j , e em_j é o erro médio do comprimento j .

A tabela mostra que a maior porcentagem de erro advém de linhas com comprimentos abaixo de 500 m para ambos os cenários, i.e., quanto menor o comprimento do cabo, pior é a estimativa de comprimento. Isto mostra que para distâncias pequenas, a

Tabela 17 – Comprimentos estimados e erro de todas as 164 medições S_{11}^{PM} .

Real (m)	Estimado (m)	Erro (%)	Real (m)	Estimado (m)	Erro (%)	Real (m)	Estimado (m)	Erro (%)
200	283,71	41,85	500	561,39	12,28	150	172,74	15,16
200	283,71	41,85	500	539,8	7,96	150	172,74	15,16
200	276,04	38,02	500	575,79	15,16	150	179,93	19,96
200	283,71	41,85	500	561,39	12,28	150	172,74	15,16
200	276,04	38,02	500	547	9,4	150	172,74	15,16
200	210,55	5,27	500	547	9,4	150	172,74	15,16
250	337,38	34,95	500	511,01	2,2	150	172,74	15,16
250	337,38	34,95	500	568,59	13,72	200	233,94	16,97
250	329,71	31,89	500	518,21	3,64	200	233,94	16,97
250	329,71	31,89	500	511,01	2,2	200	237,51	18,76
250	322,05	28,82	500	568,59	13,72	200	237,51	18,76
500	497,75	0,45	500	532,61	6,52	250	287,89	15,16
500	495,1	0,98	500	547	9,4	250	287,89	15,16
500	516,28	3,26	500	525,41	5,08	250	295,09	18,04
500	571,88	14,38	500	539,8	7,96	250	287,89	15,16
500	532,17	6,43	500	568,59	13,72	250	295,09	18,04
500	496,62	0,68	500	568,59	13,72	250	287,89	15,16
500	475,03	4,99	500	561,39	12,28	250	302,29	20,92
500	491,27	1,75	500	506,65	1,33	250	287,89	15,16
500	496,62	0,68	500	545,62	9,12	250	287,89	15,16
500	503,82	0,76	500	514,44	2,89	250	287,89	15,16
500	496,62	0,68	500	530,03	6,01	500	516,28	3,26
500	503,82	0,76	500	514,44	2,89	500	524,22	4,84
500	503,82	0,76	500	537,83	7,57	500	526,87	5,37
500	482,22	3,56	100	122,36	22,36	500	529,52	5,9
500	499,07	0,19	100	115,16	15,16	500	529,52	5,9
500	499,07	0,19	100	115,16	15,16	500	534,81	6,96
500	499,07	0,19	100	115,16	15,16	500	516,28	3,26
500	496,62	0,68	100	115,16	15,16	500	550,7	10,14
500	489,42	2,12	100	122,36	22,36	500	500,39	0,08
500	489,42	2,12	100	115,16	15,16	500	500,39	0,08
500	499,07	0,19	100	115,16	15,16	500	500,39	0,08
500	499,07	0,19	100	122,36	22,36	500	532,61	6,52
500	475,68	4,86	100	115,16	15,16	500	582,99	16,6
500	471,27	5,75	100	115,16	15,16	500	539,8	7,96
500	497,75	0,45	100	115,16	15,16	500	575,79	15,16
500	497,75	0,45	100	122,36	22,36	500	525,41	5,08
500	473,92	5,22	100	115,16	15,16	500	575,79	15,16
500	539,8	7,96	100	115,16	15,16	500	547	9,4
500	518,21	3,64	150	172,74	15,16	500	547	9,4
500	511,01	2,2	150	179,93	19,96	500	568,59	13,72
500	582,99	16,6	150	165,54	10,36	500	532,61	6,52
500	561,39	12,28	150	172,74	15,16	500	561,39	12,28
500	547	9,4	150	158,34	5,56	500	561,39	12,28
500	525,41	5,08	150	172,74	15,16	500	539,8	7,96
500	554,2	10,84	150	172,74	15,16	500	539,8	7,96
500	561,39	12,28	150	172,74	15,16	500	539,8	7,96
500	554,2	10,84	150	165,54	10,36	500	547	9,4
500	532,61	6,52	150	172,74	15,16	500	532,61	6,52
500	539,8	7,96	150	158,34	5,56	500	518,21	3,64
500	547	9,4	150	172,74	15,16	500	518,21	3,64
500	539,8	7,96	150	179,93	19,96	500	518,21	3,64
500	547	9,4	150	179,93	19,96			
500	532,61	6,52	150	165,54	10,36			
500	561,39	12,28	150	172,74	15,16			
500	518,21	3,64	150	172,74	15,16			

técnica de estimativa de comprimento aparenta ligeira confusão e admiti erros maiores de estimativa para ambos os cenários.

O gráfico em barras (Figura 41) mostra informações estatísticas a cerca do método de estimativa de comprimento. Este informação é obtida através do intervalo de confiança de 95% para cada um dos três cenário mostrado: mesmo *binder*, *binders* diferentes e total.

Tabela 18 – Distribuição de erro (%)

	MB	BD	Total
Média	9,84	12,7	11,3
Desv. padrão	11,14	5,61	8,84

Tabela 19 – Erro por intervalo (%)

Faixa	Erro	MB	BD
[0,10]	49,39	68,75	30,95
]10,20]	41,46	18,75	63,1
]20,30]	3,66	1,25	5,95
]30,40]	3,66	7,5	0
]40,50]	1,83	3,75	0
]50,100]	0	0	0

Tabela 20 – Erro médio proporcional separado por comprimento.

Tipo	100 m	150 m	200 m	250 m	500 m
Mesmo <i>binder</i>	-	-	47,21	44,5	8,29
<i>Binders</i> diferentes	23,38	19,93	24,46	22,33	10,01

A margem de erro para cada cenário é calculada através da equação seguinte

$$me = Z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}} \quad (5.5)$$

onde $Z_{\alpha/2}$ é o valor consultado na tabela Z para o intervalo de confiança α , σ é o desvio padrão e n é o número de amostras. Para o intervalo de confiança de 95%, tem-se o valor correspondente de 1,96 extraído da tabela Z . O resultado da equação é a margem de erro me associada ao intervalo de confiança escolhido.

A margem de erro para o cenário de mesmo *binder* foi de 2,44%. Já para o cenário de *binders* diferentes a margem de erro foi de 1,2%. E ainda, 1,35% considerando todas as medições. Devido ao elevado valor de desvio padrão para o cenário de mesmo *binder*, obteve-se um alto valor para a margem de erro, no entanto, em média, mantém-se ainda abaixo do teto de todos os outros cenários.

Pode-se concluir que a técnica proposta de estimação de comprimento mostrou-se melhor no cenário de *mesmo binder* onde as medições se assemelham à medições de sinais em modo diferencial em pares individuais. O maior acoplamento devido a proximidade dos PTs e menor descasamento de impedância com o equipamento de medição, contruí significativamente para que boa parte do sinal fantasma seja inserido no circuito. Consequentemente, a reflexão do sinal no domínio do tempo referente a PTs que estão no *binder* e se separam em *binders* diferentes é mais perceptível. Quanto mais definido

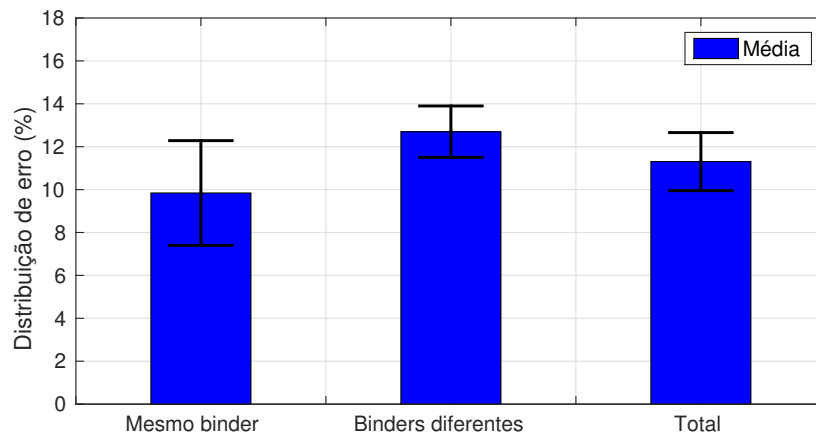


Figura 41 – Gráfico em barras para os três cenários distintos. Margem de erro em diferentes visões: mesmo *binder* (2,44%), *binders* diferentes (1,2%) e todas as medições (1,35%). O intervalo de confiança usado foi de 95%.

for este pulso refletido, melhor será a detecção de singularidades lógicas, i.e., que façam sentido e remetam a reflexão correta referente ao final do *binder*.

No outro cenário, *binders* diferentes, a técnica apresentou dificuldades em estimar os comprimentos dos *binders*. O baixo acoplamento entre os PTs provocado pela maior distância entre eles e maior descasamento de impedância com o equipamento de medição, contribuí para que parte do sinal seja refletido na entrada do circuito, conseqüentemente a outra parte do sinal que é injetada no canal possui menos energia, de tal forma que a reflexão do sinal referente ao ponto em que os PTs se dividem é menos perceptível. Isto dificulta a detecção de singularidades lógicas no domínio do tempo.

Conclusão

Este trabalho desenvolveu um método geral para a identificação automática de *binders* e estimativa de comprimento para pares trançados situados no mesmo cabo. Os algoritmos de aprendizado de máquina utilizados na identificação de *binders* têm abordagem supervisionada (SVM) e não supervisionada (*K*-means e GMM). O algoritmo SVM requer a construção de dois classificadores, uma para a identificação de cabos e outro para a identificação de *binders*. Este procedimento é necessário visto que o SVM resolve problemas binários pela distinção de amostras oriundas de dois padrões através de hiperplanos. Já os algoritmos de clusterização conseguem reconhecer padrões vários padrões distintos através de um único classificador.

Um método de rotulação automática de *clusters* foi desenvolvido com o objetivo de identificar o significado de cada *cluster* encontrado pelos algoritmos *K*-means e GMM. Este método torna possível calcular a acurácia de um modelo construído pelos algoritmos de clusterização. O método desenvolvido mostrou ser coerente, visto que embarca o conhecimento necessário para reconhecer os três padrões trabalhos nesta dissertação e permite calcular a acurácia dos modelos produzidos através da abordagem não supervisionada. A análise de características foi um recurso fundamental para estudar o domínio do problema e permitiu descobrir quais características extraídas do sinal são mais revelantes, contribuindo para que o desenvolvimento do método de rotulação de *clusters* utilizasse tão somente informações importantes para realizar a identificação dos padrões.

Sobre análise geral dos algoritmos de aprendizado de máquina aplicados neste trabalho, pode-se afirmar que a identificação de cabos é uma tarefa mais fácil se comparada a identificação de *binder*. Isto provavelmente ocorre porque na identificação de cabo, distância e material entre os condutores afeta fortemente a impedância característica, criando uma assinatura bem definida nas medições e gerando forte influência na tarefa de classificação em relação a identificação de *binders*. Revela-se também que as medições fantasmas feitas em cabos blindados são mais fáceis de se identificar, pois os campos elétricos e magnéticos estão mais confinados dentro do cabo, evitando o acoplamento entre dois PTs que estão em cabos diferentes, fazendo com que a medição S_{11}^{PM} perca quase toda sua periodicidade e aumente o nível de efeito NER. Isto favorece também a identificação de *binder*, uma vez que com campos mais confinados, menor é a interferência externa, aumentando o acoplamento entre dois PTs dentro de uma cabo e intensificando a periodicidade do sinal S_{11}^{PM} .

Na identificação de *binder* em cabos não blindados, principalmente quando dois PTs estão no mesmo *binder*, i.e., estão próximos uns dos outros, distância entre condutores não é

mais tão relevante e outras características geométricas e físicas dos cabos devem influenciar nas medições, tais como taxa de trançamento entre os pares, taxa de trançamento entre os *binders*, não homogeneidade do meio e não uniformidade presentes nos condutores. Estas características influenciam na classificação de *binders* e, conseqüentemente, nos resultados alcançados pelos algoritmos de aprendizado de máquina, penalizando principalmente a classificação correta das amostras de PTs que estão no mesmo *binder*.

O método para identificação de *binders*, utilizando os algoritmos SVM e *K*-means, obteve 100% de acurácia quando o propósito é identificar se dois PTs estão em cabos diferentes, com GMM alcançando 97,75% de acurácia. O algoritmo SVM alcançou 98,88% de acurácia geral, ultrapassando *K*-means com 91,76% e GMM com 92,13%. O algoritmo SVM também superou o *K*-means e GMM na identificação de *binders*. Porém, os algoritmos não supervisionados usados neste trabalho são mais simples e não necessitam que as amostras da base de treinamento tenham suas classes predefinidas, já o SVM funciona mediante esta classificação prévia.

Em relação às técnicas não supervisionadas, GMM superou o algoritmo *K*-means na identificação automática de *binders*. GMM é uma técnica de clusterização suave, o que implica dizer que uma amostra é classificada em relação a uma classe com um certo grau de certeza (probabilidade posterior). Por outro lado, *K*-means é uma técnica de clusterização rígida, i.e., assume que uma amostra pertence ou não pertence totalmente a um *cluster*. Neste contexto, o algoritmo GMM se torna ainda mais interessante de ser usado, pois passa ao operador a possibilidade, em última instância, de decidir corretamente se uma amostra pertence ou não a um determinado cabo.

A análise nos resultados da estimativa de comprimento mostrou que medições S_{11}^{PM} em *binders* diferentes são mais difíceis de estimar o comprimento do que medições S_{11}^{PM} que estão no mesmo *binder*. O baixo acoplamento entre PTs em *binders* diferentes diminui a qualidade da estimativa de comprimento. O erro médio proporcional mostrou que a técnica erra mais quando o comprimento real dos cabos diminui (< 500 m). Ainda assim, o erro médio total se manteve abaixo de 10% considerando ambos os cenários avaliados.

Como trabalhos futuros, o método de identificação automática de *binders* pode ser avaliado em redes de *binders* reais considerando os três cenários trabalhos nesta dissertação. Além disso, outras técnicas de aprendizado de máquina podem ser exploradas no problema, como por exemplo os algoritmos de clusterização *Linkage*, DBSCAN e U*C.

Publicações

- Reginaldo Santos, Claudomiro Sales Jr., Manoel Lima, Caio Rodrigues, Alessandra Araújo, Antoni Fertner, João C. W. A. Costa. “*Clustering Strategies for Binder Identification using Phantom Measurements*”. IEEE Global Communications Conference, Exhibition & Industry Forum - Globecom, 2015.
- Reginaldo Santos, Claudomiro Sales, Manoel Lima, Caio Rodrigues, Alessandra Araújo, Walisson Cardoso, Antoni Fertner, João C. W. A. Costa. “*Binder Identification using Pattern Recognition on Phantom Measurements*”. IEEE Transactions on Instrumentation and Measurement, 2015.

Referências

- 1 GAMBINI, J.; SPAGNOLINI, U. Wireless over cable for femtocell systems. *Communications Magazine, IEEE*, v. 51, n. 5, p. 178–185, May 2013. ISSN 0163-6804. Citado 5 vezes nas páginas [9](#), [15](#), [17](#), [20](#) e [32](#).
- 2 COOMANS, W. et al. The 5th generation broadband copper access. In: *Broadband Coverage in Germany. 9th ITG Symposium. Proceedings*. [S.l.: s.n.], 2015. p. 1–5. Citado 4 vezes nas páginas [9](#), [15](#), [21](#) e [32](#).
- 3 G.9700, I.-T. *Fast Access to Subscriber Terminals (G.fast) - Power spectral density specification*. [S.l.], 2014. Citado na página [15](#).
- 4 G.9701, I.-T. *Fast Access to Subscriber Terminals (G.fast) - Physical layer specification*. [S.l.], 2014. Citado na página [15](#).
- 5 TIMMERS, M. et al. G.fast: evolving the copper access network. *Communications Magazine, IEEE*, v. 51, n. 8, p. –, 2013. ISSN 0163-6804. Citado 3 vezes nas páginas [15](#), [19](#) e [32](#).
- 6 WEI, D. et al. G.fast for ftt dp: Enabling gigabit copper access. In: *Globecom Workshops (GC Wkshps), 2014*. [S.l.: s.n.], 2014. p. 668–673. Citado na página [15](#).
- 7 MEDEIROS, E. et al. How vectoring in g.fast may cause neighborhood wars. In: *Communications (ICC), 2014 IEEE International Conference on*. [S.l.: s.n.], 2014. p. 3859–3864. Citado na página [15](#).
- 8 TIMMERS, M. et al. System design of reverse-powered g.fast. In: *Communications (ICC), 2012 IEEE International Conference on*. [S.l.: s.n.], 2012. p. 6869–6873. ISSN 1550-3607. Citado na página [15](#).
- 9 COOMANS, W. et al. Xg-fast: Towards 10 gb/s copper access. In: *Globecom Workshops (GC Wkshps), 2014*. [S.l.: s.n.], 2014. p. 630–635. Citado 4 vezes nas páginas [15](#), [19](#), [21](#) e [32](#).
- 10 GAMBINI, J. et al. Wireless over cable in femtocell systems: A case study from indoor channel measurements. In: *Wireless Communications and Networking Conference Workshops (WCNCW), 2012 IEEE*. [S.l.: s.n.], 2012. p. 137–141. Citado na página [15](#).
- 11 LEUNG, C. et al. Vectored dsl: Potential, implementation issues and challenges. *Communications Surveys Tutorials, IEEE*, v. 15, n. 4, p. 1907–1923, 2013. ISSN 1553-877X. Citado 4 vezes nas páginas [15](#), [17](#), [18](#) e [19](#).
- 12 ZIDANE, R. et al. Vectored dsl: benefits and challenges for service providers. *Communications Magazine, IEEE*, v. 51, n. 2, p. 152–157, February 2013. ISSN 0163-6804. Citado na página [15](#).
- 13 MAES, J.; NUZMAN, C. Energy efficient discontinuous operation in vectored g.fast. In: *Communications (ICC), 2014 IEEE International Conference on*. [S.l.: s.n.], 2014. p. 3854–3858. Citado na página [15](#).

- 14 GOLDEN, P.; DEDIEU, H.; JACOBSEN, K. *Fundamentals of DSL Technology*. [S.l.]: Auerbach Publications, 2006. Citado na página 16.
- 15 NEUS, C. et al. Binder identification by means of phantom measurements. *Instrumentation and Measurement, IEEE Transactions on*, v. 60, n. 6, p. 1967–1975, 2011. ISSN 0018-9456. Citado 6 vezes nas páginas 16, 18, 22, 46, 48 e 53.
- 16 SALES, C. et al. Line topology identification using multi-objective evolutionary computation. *IEEE Transactions on Instrumentation & Measurement*, v. 59, n. 3, p. 715–729, mar. 2010. Citado 2 vezes nas páginas 16 e 31.
- 17 GALLI, S.; WARING, D. L. Loop makeup identification via single ended testing: Beyond mere loop qualification. *IEEE Journal on Selected Areas in Communications*, v. 20, n. 5, p. 923–935, jun. 2002. Citado 2 vezes nas páginas 16 e 32.
- 18 NEUS, C.; BOETS, P.; BIESEN, L. van. Transfer function estimation of digital subscriber lines with single ended line testing. In: *Proceedings of the IEEE Instrumentation and Measurement Technology Conference*. [S.l.: s.n.], 2007. p. 1–5. Citado na página 16.
- 19 FOUBERT, W. et al. Exploiting the phantom-mode signal in dsl applications. *Instrumentation and Measurement, IEEE Transactions on*, v. 61, n. 4, p. 896–902, April 2012. ISSN 0018-9456. Citado 2 vezes nas páginas 16 e 19.
- 20 LAFATA, P.; JARES, P.; VODRAZKA, J. Increasing the transmission capacity of digital subscriber lines. In: *Telecommunications and Signal Processing (TSP), 2012 35th International Conference on*. [S.l.: s.n.], 2012. p. 292–296. Citado 2 vezes nas páginas 18 e 28.
- 21 PEETERS, M.; VANHASTEL, S. *The Copper Phantom*. 2013. OSP Magazine. Disponível em: <<http://www.ospmag.com/issue/article/The-Copper-Phantom>>. Citado na página 18.
- 22 WILLIAMS, J. *How Enhanced DSL Technologies Optimize the Last Copper Mile*. 2013. JDS Uniphase Corporation. Disponível em: <<http://www.jdsu.com/ProductLiterature/dsltech-wp-tfs-tm-ae.pdf>>. Citado 2 vezes nas páginas 18 e 19.
- 23 LEE, B. et al. Binder mimo channels. *Communications, IEEE Transactions on*, v. 55, n. 8, p. 1617–1628, Aug 2007. ISSN 0090-6778. Citado na página 18.
- 24 FOUBERT, W. et al. Exploiting the phantom-mode signal in dsl applications. *Instrumentation and Measurement, IEEE Transactions on*, v. 61, n. 4, p. 896–902, 2012. ISSN 0018-9456. Citado 3 vezes nas páginas 18, 28 e 30.
- 25 GOMES, D. de A. *Transmissão DSL em Modo Fantasma: Medições e Avaliação de Desempenho*. 2012. Disponível em: <<http://repositorio.ufpa.br/jspui/handle/2011/3315>>. Citado 2 vezes nas páginas 18 e 29.
- 26 RHEE, W. et al. *Binder identification*. Google Patents, 2006. WO Patent App. PCT/IB2006/000,836. Disponível em: <<http://www.google.com/patents/WO2006120513A1?cl=en>>. Citado na página 18.
- 27 G.998.1, I.-T. *ATM-based multi-pair bonding*. [S.l.], 2005. Citado na página 19.
- 28 G.998.2, I.-T. *Ethernet-based multi-pair bonding*. [S.l.], 2005. Citado na página 19.

- 29 G.998.3, I.-T. *Multi-pair bonding using time-division inverse multiplexing*. [S.l.], 2005. Citado na página 19.
- 30 HINCAPIE, D. et al. Evaluation of binder management for partially controlled dsl vectoring systems. In: *Communications (ICC), 2015 IEEE International Conference on*. [S.l.: s.n.], 2015. p. 964–970. Citado na página 19.
- 31 LAFATA, P. Estimations of g.fast transmission performance over phantom modes. In: *Telecommunications and Signal Processing (TSP), 2015 38th International Conference on*. [S.l.: s.n.], 2015. p. 1–5. Citado 2 vezes nas páginas 19 e 22.
- 32 KERPEZ, K. et al. The impact of plc-to-dsl interference on vdsl2, vectored vdsl2, and g.fast. In: *Power Line Communications and its Applications (ISPLC), 2015 International Symposium on*. [S.l.: s.n.], 2015. p. 160–165. Citado na página 19.
- 33 ALI, K.; MESSIER, G.; LAI, S. Dsl and plc co-existence: An interference cancellation approach. *Communications, IEEE Transactions on*, v. 62, n. 9, p. 3336–3350, Sept 2014. ISSN 0090-6778. Citado na página 19.
- 34 RADCLIFFE, D. *Who's the world's fibre broadband leader?* 2014. Disponível em: <<http://www.zdnet.com/article/whos-the-worlds-fibre-broadband-leader-prepare-to-be-surprised/>>. Citado na página 21.
- 35 ODLING, P. et al. The fourth generation broadband concept. *IEEE Communications Magazine*, v. 47, n. 1, p. 62–69, jan 2009. ISSN 0163-6804. Citado na página 26.
- 36 GALLI, S.; KERPEZ, K. J. Single-ended Loop Make-up Identification-Part I: a Method of Analyzing TDR Measurements. *IEEE Transactions on Instrumentation and Measurement*, v. 55, n. 2, p. 528–537, abr. 2006. Citado na página 32.
- 37 LONG, G.; KAMALI, J. Single-ended line probing helps speed up DSL mass deployment. In: *IIC-China/ESC China Conference*. [S.l.: s.n.], 2002. p. 57–60. Citado na página 32.
- 38 CHEN, C.-J. L. P.-H.; SCHÖLKOPF, B. *A Tutorial on v-Support Vector Machines*. [S.l.], 2013. Citado na página 34.
- 39 MULLER, K. et al. An introduction to kernel-based learning algorithms. *Neural Networks, IEEE Transactions on*, v. 12, n. 2, p. 181–201, 2001. ISSN 1045-9227. Citado na página 34.
- 40 HOFMANN, T.; SCHÖLKOPF, B.; SMOLA, A. J. Kernel methods in machine learning. *The annals of statistics*, JSTOR, p. 1171–1220, 2008. Citado na página 36.
- 41 ALSABTI, K.; RANKA, S.; SINGH, V. An efficient k-means clustering algorithm. *Proc. First Workshop High Performance Data Mining, Mar. 1988*, 1988. Citado na página 38.
- 42 DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, v. 39, n. 1, p. 1–38, 1977. Citado na página 40.

- 43 WITTEN, I. H.; FRANK, E. *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems)*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2005. ISBN 0120884070. Citado 2 vezes nas páginas 43 e 44.
- 44 KIM, K. et al. Cyclostationary approaches to signal detection and classification in cognitive radio. In: *New Frontiers in Dynamic Spectrum Access Networks, 2007. DySPAN 2007. 2nd IEEE International Symposium on*. [S.l.: s.n.], 2007. p. 212–215. Citado na página 50.
- 45 AB, E. *Access Network Pair cable, TEL 313 000*. [S.l.], 2010. Citado na página 52.
- 46 AB, E. *HF Pair cables TEL 481 02*. [S.l.], 2010. Citado na página 52.
- 47 LIMA, V. D. et al. A wavelet-based expert system for dsl line topology identification. *International Journal of Communication Systems*, 2014. Citado 2 vezes nas páginas 62 e 63.
- 48 SALES, C. et al. Expert system based on wavelets and delat measurements for vdsl systems. In: *Global Communications Conference (GLOBECOM), 2012 IEEE*. [S.l.: s.n.], 2012. p. 3110–3115. ISSN 1930-529X. Citado 2 vezes nas páginas 62 e 63.
- 49 GALLI, S.; KERPEZ, K. Single-ended loop make-up identification-part i: a method of analyzing tdr measurements. *Instrumentation and Measurement, IEEE Transactions on*, v. 55, n. 2, p. 528–537, April 2006. ISSN 0018-9456. Citado na página 62.