UNIVERSIDADE FEDERAL DE CAMPINA GRANDE

CENTRO DE ENGENHARIA ELÉTRICA E INFORMÁTICA

UNIDADE ACADÊMICA DE SISTEMAS E COMPUTAÇÃO

COORDENAÇÃO DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

MIQUEAS GALDINO DOS SANTOS

ABORDAGEM PARA CATEGORIZAÇÃO DE ANOMALIAS EM REDES DE SENSORES SEM FIO BASEADO EM LÓGICA FUZZY

CAMPINA GRANDE - PB

2021

Universidade Federal de Campina Grande
Centro de Engenharia Elétrica e Informática
Coordenação de Pós-Graduação em Ciência da Computação

Abordagem para Categorização de Anomalias em Redes de Sensores Sem Fio baseado em Lógica Fuzzy

Miqueas Galdino dos Santos

Dissertação submetida à Coordenação do Curso de Pós-Graduação em Ciência da Computação da Universidade Federal de Campina Grande - Campus I como parte dos requisitos necessários para obtenção do grau de Mestre em Ciência da Computação.

Área de Concentração: Ciência da Computação
Linha de Pesquisa: Redes de Computadores

Reinaldo Cézar de Morais Gomes
(Orientador)
Ruan Delgado Gomes
(Coorientador)

Campina Grande, Paraíba, Brasil

S237a Santos, Miqueas Galdino dos.
Abordagem para categorização de anomalias em Redes de Sensores sem Fio baseado em Lógica Fuzzy / Miqueas Galdino dos Santos. – Campina Grande, 2021.
81 f. : il. color.

Dissertação (Mestrado em Ciência da Computação) – Universidade Federal de Campina Grande, Centro de Engenharia Elétrica e Informática, 2021.
"Orientação: Prof. Dr. Reinaldo Cézar de Morais Gomes; Coorientação: Prof. Dr. Ruan Delgado Gomes".
Referências.

1. Redes Sensores sem Fio (RSSFs). 2. Internet das Coisas. 3. Detecção de Anomalias. 4. Categorização de Anomalias. I. Gomes, Reinaldo Cézar de Morais. II. Gomes, Ruan Delgado. III. Título.

CDU 004.738:621.316 (043)

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECÁRIA MARIA ANTONIA DE SOUSA CRB 15/398

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
POS-GRADUACAO CIENCIAS DA COMPUTACAO
Rua Aprigio Veloso, 882, - Bairro Universitario, Campina Grande/PB, CEP 58429-900

FOLHA DE ASSINATURA PARA TESES E DISSERTAÇÕES

MIQUEAS GALDINO DOS SANTOS

ABORDAGEM PARA CATEGORIZAÇÃO DE ANOMALIAS EM REDES DE SENSORES SEM FIO BASEADO EM LÓGICA FUZZY

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação como pré-requisito para obtenção do título de Mestre em Ciência da Computação.

Aprovada em: 19/08/2021

Prof. Dr. REINALDO CÉZAR DE MORAIS GOMES, Orientador, UFCG

Prof. Dr. RUAN DELGADO GOMES, Orientador, IFPB

Prof. Dr. ANDERSON FABIANO BATISTA FERREIRA DA COSTA, Examinador Interno, IFPB

Prof. Dr. IGUATEMI EDUARDO DA FONSECA, Examinador Externo, UFPB

Documento assinado eletronicamente por **REINALDO CEZAR DE MORAIS GOMES**, **PROFESSOR(A) DO MAGISTERIO SUPERIOR**, em 22/09/2021, às 15:25, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

Documento assinado eletronicamente por **Iguatemi Eduardo da Fonseca**, **Usuário Externo**, em 23/09/2021, às 09:35, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

Documento assinado eletronicamente por **Ruan Delgado Gomes**, **Usuário Externo**, em 23/09/2021, às 10:51, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

Documento assinado eletronicamente por **Anderson Fabiano Batista Ferreira da Costa**, **Usuário Externo**, em 23/09/2021, às 11:02, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

A autenticidade deste documento pode ser conferida no site https://sei.ufcg.edu.br/autenticidade, informando o código verificador **1790828** e o código CRC **3F947A59**.

Referência: Processo nº 23096.050379/2021-15 SEI nº 1790828

Resumo

Os avanços na microeletrônica permitiram a ascensão das Redes Sensores sem fio (RSSFs), que estão cada vez mais presentes em nosso dia-a-dia como um elemento fundamental para o paradigma da Internet das Coisas. Neste ambiente, a confiabilidade dos dados que transitam nessa rede é um fator relevante que gera investigações e pesquisas no ambiente acadêmico. Devido às diversas limitações existentes na arquitetura das RSSFs, falhas de sensores são comuns gerando dados incongruentes e anormais. Porém, anormalidades também refletem alterações do fenômeno que está sendo monitorado pelos sensores, gerando assim problemas na definição do que realmente está acontecendo em um determinado sensor. Assim, anomalias são indicativos de que algo fora do padrão ocorre na rede, e saber a causa dessas anormalidades é de essencial importância para tomadas de decisões no ambiente. Tendo em vista este contexto, o presente trabalho desenvolve uma abordagem de detecção e categorização de anomalias em redes de sensores sem fio baseado em lógica fuzzy, que tem por objetivo auxiliar na determinação da existência de eventos ou de sensores falhos. Sendo avaliados contextos de diferentes tipos de falhas nos dados e qual sua relação com fatores ligados a quantidade de sensores falhos numa região e perda de pacotes. Os resultados apontaram para a efetividade na identificação das anormalidades e categorização de anomalias, possuindo maior eficácia na categorização de falhas intermitentes, em relação a anomalias graduais e eventos. Também se constatou maior efetividade para ambientes com menos sensores falhos e se percebeu uma relação moderada em relação a abordagem e a perda de pacotes do ambiente.

Abstract

Advances in microelectronics have allowed the rise of Wireless Sensor Networks (WSNs), which are increasingly present in our daily lives as a fundamental element of the Internet of Things paradigm. In this environment, the reliability of the data that transits this network is a relevant factor that generates investigations and research in the academic environment. Due to the several limitations existing in the WSNs architecture, sensor failures are common, generating incongruous and abnormal data. However, abnormalities also reflect changes in the phenomenon being monitored by the sensors, thus creating problems in defining what is really happening in a given sensor. Thus, anomalies are indicative that something non-standard occurs in the network, and knowing the cause of these abnormalities is essential for decision-making in the environment. In view of this context, the present work develops an approach for detecting and categorizing anomalies in wireless sensor networks based on fuzzy logic, which aims to help determine the existence of events or faulty sensors. Contexts of different types of data failures were evaluated and what is their relationship with factors related to the number of failed sensors in a region and packet loss. The results pointed to the effectiveness in the identification of abnormalities and categorization of anomalies, with greater effectiveness in the categorization of intermittent failures, in relation to gradual anomalies and events. It was also found greater effectiveness for environments with fewer faulty sensors and a moderate relationship was noticed in relation to approach and the loss of packets in the environment.

Agradecimentos

A Deus por sua infinita misericórdia e amor para comigo, por ter me auxiliado e me dado sabedoria para conseguir ir além dos meus desejos e expectativas. Soli Deo Gloria!

A minha família que sempre me apoiou e me deu liberdade para seguir em minhas escolhas, além de todo o carinho e compreensão que me deram suporte para vencer os momentos mais difíceis, todo o meu amor e gratidão.

A Gabriela Ferreira, minha esposa e meu amor, que me ajudou como nunca, demonstrando todo seu companheirismo e afeto, todo meu carinho e gratidão à você.

Aos meus orientadores, Reinaldo Gomes e Ruan Delgado, por todo o carinho, conselho, auxilio e ajuda na minha construção como aluno, pesquisador e profissional, permitindo a realização desse trabalho. Sem tais incentivos, e conselhos, tudo isso não poderia ser realizado, minha eterna gratidão.

Aos meus colegas de laboratório, Marcela, Yngrid e Eduardo, que me ajudaram nessa caminhada com carinho e companheirismo.

Aos meus amigos do IFPB que continuam a vida toda, mais conhecidos como mascotes: Rodolfo, Maxsuel, Wemerson e Nathalya. Nunca esquecerei todos os momentos e aprendizados vividos, minha eterna gratidão a todo o trajeto construído juntos.

Aos meus amigos do Semiredelu: Sergio, Renata, Deborah e Luciana, pelo carinho, apoio e risadas ao longo de muitos anos, minha eterna gratidão.

A CAPES e ao povo brasileiro pelo custeio da bolsa ao longo desta jornada. Espero poder retribuir o apoio obtido ao longo dos anos.

Minha gratidão a todos que me auxiliaram, oraram e me incentivaram a construir esse projeto. Meu muito obrigado! Amo a Todos Vocês!

Conteúdo

Lista de Símbolos

IoT - *Internet of Things* (Internet das Coisas)

MAD - *Median absolute deviation* (Desvio Absoluto da Mediana)

MZS - *Modified Z-Score* (*Z-Score* Modificado)

QoS - *Quality of Service* (Qualidade de Serviço)

RSSF - *Redes de Sensores Sem Fio*

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

A Internet das Coisas (do inglês *Internet of Things (IoT)*) é um paradigma de interconexão que vem angariando espaço no campo das telecomunicações [54]. Sendo uma extensão da Internet atual, a IoT surgiu com o objetivo de interconectar objetos do dia-a-dia com uma rede, permitindo a interação e cooperação desses objetos, proporcionando o controle de forma remota dos mesmos e viabilizando o provimento de serviços por eles [54][65].

O surgimento de processadores e sensores de baixo consumo de energia, redes sem fio inteligentes, conjuntamente à análise de *big data*, levaram a um crescente interesse do mundo tecnológico a Internet das Coisas. Diferentemente de tempos passados, a conexão com a Internet não está mais limitada aos computadores convencionais e tal conexão abrange uma grande e heterogênea quantidade de equipamentos, como: TVs, smartphones, automóveis, relógios, consoles de jogos, *web cams*, entre outros.

Um dos fatores mais relevantes e atraentes para os usuários do paradigma é, dentre suas variadas funcionalidades e características, possuir a capacidade de mensurar grandezas físicas do ambiente por intermédio dos sensores que compõem o sistema, de forma que ao se coletar os dados oriundos do ambiente, seja possível processar informações e compreender melhor os fenômenos que estão sendo analisados [53][25].

Isto posto, um dos principais arcabouços de um sistema IoT são as Redes de Sensores Sem Fio (RSSFs), que são redes caracterizadas pela utilização de nós sensores distribuídos em uma determinada localidade, apresentando a capacidade de comunicação entre si por meio de enlaces sem fio e que têm como objetivo monitorar um determinado fenômeno ou grandeza, como: temperatura, som, níveis de poluição, umidade, entre outros [68]. Isso

ocorre, devido a composição dos nós, que possuem transceptores de rádio, fonte de energia, sensores e microcontroladores. Arbitrariamente, os nós sensores, podem acrescentar módulos de localização, geração de energia e atuadores[49]. Assim, através dessas redes, dados e informações sobre determinados eventos podem ser extraídas e entregues a aplicações ou usuários.

A evolução tecnológica e o crescimento mercadológico do conceito *IoT* gera novas possibilidades de aplicações, que são empregadas em paradigmas como: Cidades Inteligentes (*Smart Cities*), Saúde (*Healthcare*), Casas inteligentes (*Smart Homes*) e a Indústria 4.0. Neste cenário, o mercado das Redes de Sensores sem fio se estabelecem, segundo dados da *Mordor Intelligence*, O tamanho do mercado global de rede de sensores sem fio era de US $46.76 bilhões em 2020 e está projetado para atingir US $ 123,93 bilhões em 2026, crescimento de mais de quase 300%. Isso tudo devido a inserção cada vez maior do conceito de IoT e a crescente necessidade de monitoramento de dados em tempo real entre organizações de variados setores como de saúde, automotivo, manufatura entre outros [1].

Identifica-se, então, que os dados são a principal fonte de geração de conhecimento sobre o mundo físico, trazendo informações sobre todo o processo monitorado, desde a análise comportamental do fenômeno analisado até possíveis erros relacionados a sensores ou variações de enlace [25]. Os dados então representam um ativo valioso nas RSSFs, e consequentemente na IoT. Isso porque, através deles é que as informações sobre um determinado fenômeno, pessoa ou entidade podem ser geradas e posteriormente utilizadas na construção de serviços inteligentes [13]. Enfatizando assim a importância do conceito de qualidade de dados para as RSSFs, conceito este que aponta que para que a informação seja segura e confiável os dados necessitam de validade, autenticidade e confiabilidade [21].

Um dos obstáculos que surgem afetando a qualidade de dados e a confiabilidade do monitoramento das RSSF são as chamadas anomalias. As anomalias são irregularidades que podem ser causadas por diversos fatores, dentre eles: falhas de *hardware* dos sensores, ataques aos dispositivos ou alterações na variável de monitoramento [52].

Essas anomalias geralmente são apontadas e averiguadas através de ferramentas de gerenciamento de rede, porém tais métodos não são necessariamente acurados. Em alguns casos, uma simples verificação nos níveis das medidas é o bastante para verificar sua validade, entretanto, outros desvios de padrão dos dados de tráfego podem ser sutis e de difícil

detecção em um simples monitoramento de níveis de sinais, de forma que tais desvios podem gerar alterações nas distribuições de probabilidade temporal e espacial. Além disso, tais tipos de ferramentas identificam a existência de anomalias, mas não conseguem classifica-las ou categoriza-las de acordo com a provável causa de surgimento das mesmas.

1.1 Formulação do Problema

A utilização das redes sensores sem fio possui vantagens em relação ao uso de redes cabeadas. Pouca flexibilidade de cabeamento, dificultando a instalação e manutenção da rede, gerando impactos inclusive no custo de implementação, em que geralmente em uma rede cabeada tal processo é mais oneroso do que os próprios sensores, são gargalos das redes cabeadas que abrem espaço as RSSFs [35][46]. As RSSF por consequente surgem com um menor custo e com uma capacidade de flexibilização e auto-organização muito maior, além da capacidade de implantação em locais mais hostis [19][46].

Assim, o uso da comunicação sem fio gera determinados benefícios, mas faz com que vários outros problemas também surjam. A alta atenuação e o surgimento de maiores obstruções do ambiente, são alguns deles. Esse impasses apresentam-se de modos diferentes, como perda de dados, falhas nos dados devido a atrasos de transmissão e *jitters* de amostragem[13].

Um dos problemas no funcionamento das RSSFs são as aparições de anomalias nos dados sensoreados, oriundos de falhas de confiabilidade, de nós sensores defeituosos ou fenômenos incomuns na zona de monitoramento, os chamados eventos [48]. No mundo real, anomalias oriundas de falhas de nós isolados podem derrubar toda a rede, trazendo sérios danos ao sistema, o que é prejudicial à confiabilidade das RSSFs [66], enquanto que eventos apontam a existência de acontecimentos e fenômenos na região monitorada, sendo de suma importância sua identificação.

Dadas as múltiplas causas de surgimento de anomalias, alguns problemas despontam para o desenvolvimento de mecanismos e protocolos de monitoramento de dados no ambiente de redes sensores sem fio. Dentre eles, pode-se citar:

- Como um nó sorvedouro (ou estação base) pode distinguir uma anomalia causada pelo fenômeno monitorado (evento) e uma anomalia causada por falha do nó sensor?

- O aumento de nós sensores considerados falhos influencia na categorização de anomalias?

- Existe algum efeito na relação entre perda de dados e identificação de origem de anomalia?

Para tentar suprimir essas limitações, o presente trabalho visa apresentar uma abordagem de detecção e categorização de anomalias, que seja capaz de auxiliar na identificação de prováveis nós defeituosos ou eventos, e a partir disto facilitar a discussão e a criação de medidas de atuação eficazes a depender da situação verificada.

1.2 Justificativa e Relevância

Tratando-se do mercado, empresas e indústrias encaram uma alta demanda para a melhora e eficiência de seus processos, de forma a atender os objetivos financeiros, corporativos e cumprir as normas ambientais. Dado a constante movimentação e evolução do mercado, são necessários sistemas inteligentes e com menores custos, para uma maior produtividade e rendimento dos empreendimentos almejados [19].

Uma RSSF comum é composta por uma determinada quantidade de nós formada por sensores que possuem restrições de recursos, com baixa capacidade de processamento e restrições de consumo de energia [14]. Habitualmente, os sensores são implantados em ambientes não controlados ou em áreas monitoradas hostis, para monitorar parâmetros críticos como vibração, temperatura, pressão e eficiência de motores, fazendo com que os nós sensores sejam vulneráveis e menos confiáveis. Esse ambiente propicia uma maior facilidade no surgimento de nós falhos, falhas essas que podem ser causadas por danos de *hardware* dos nós e/ou falhas paramétricas, que são oriundas de detecção incorretas devido a dados imprecisos coletados pelos sensores [45]. Isso tudo podendo gerar anomalias no tráfego produzido por esses nós, e consequentemente medições falhas, induzindo interpretações errôneas dos fenômenos sensoreados pela rede e dificultando a identificação de eventos [48][40].

Desta forma, as RSSFs possuem diversos desafios relacionados à confiabilidade e robustez de seus dados, de forma a executar e atingir adequadamente os objetivos de suas variadas aplicações. Assim é evidente que o estudo e desenvolvimento de uma abordagem de moni-

toramento e avaliação de qualidade de dados em uma RSSF que não só permita a detecção de dados anômalos, mas que consiga classificar a causa dessa anormalidade, seja ela oriunda de falha de sensor ou alteração na grandeza monitorada, é de total relevância para a garantia da qualidade dos dados nos sistemas desenvolvidos.

1.3 Objetivos

1.3.1 Objetivo Geral

O presente trabalho tem por objetivo desenvolver uma abordagem de detecção e categorização de anomalias para ambientes online em redes de sensores sem fio, visando identificar e classificar ocorrências anômalas ligadas a falhas sistêmicas ou individuais de sensores, em relação a eventos associados a variações contextuais do fenômeno monitorado. Assim, auxiliando na manutenção da confiabilidade da rede e otimizando ações dos monitores do sistema de redes.

1.3.2 Objetivos Específicos

- Identificar as principais causas de anomalias recorrentes em redes de sensores sem fio, suas características e causas principais;
- Estudar e definir as técnicas de detecção de anomalias que se adéquem ao ambiente considerado no estudo;
- Estudar e identificar quais as maiores limitações no desenvolvimento de mecanismos de categorização de anomalias em Redes Sensores Sem Fio;
- Criar uma abordagem de identificação e categorização de anomalias online para RSSFs;
- Desenvolver um estudo para teste e análise da eficiência da abordagem proposta.

1.4 Contribuições

Dentro do contexto das Redes de Sensores sem fio, construímos uma abordagem de categorização de anomalias online. A partir disso, caracterizamos e direcionamos a rede a identificar eventos e/ou nós falhos. Como resultado, as contribuições desta pesquisa são:

- Conseguir categorizar dados anormais em contextos de eventos e falhas de sensores;
- Conseguir analisar diversos contextos e tipos de falhas de dados;
- Construir um mecanismo de geração de anomalias artificiais, que permita a criação de ambientes distintos;
- Considerar ambientes com uma alta taxa de nós falhos.

1.5 Organização do Trabalho

A organização do trabalho se dará da seguinte forma: no Capítulo 2, são apresentados os conceitos relacionados às Redes de Sensores Sem Fio, Detecção de anomalias e anomalias no contexto das redes de sensores sem fio. No Capítulo 3 são discutidos os trabalhos relacionados a presente proposta. No Capítulo 4 se é detalhada a proposta da abordagem. No Capítulo 5 é apresentado como se deu o processo experimental para avaliação da proposta. No Capítulo 6 são apresentado os resultados referentes aos experimentos realizados e a discussão em torno destes resultados. Por fim, o Capítulo 7 apresenta as considerações finais e os trabalhos futuros relacionados a esta proposta.

Capítulo 2

Fundamentação Teórica

Este capítulo apresenta os conceitos necessários para a compreensão do trabalho. Na Seção 2.1 é fornecida uma descrição geral do funcionamento das redes de sensores sem fio. Na Seção 2.2 são apresentados os conceitos fundamentais sobre qualidade de serviço em redes sensores sem fio. Na Seção 2.3 são apresentadas as principais características da detecção de anomalias em Redes de Sensores Sem fio. Na Seção 2.4 é apresentado o conceito de lógica e sistema fuzzy.

2.1 Redes de Sensores sem Fio

Uma Rede de Sensores Sem Fio (RSSF) pode ser definida como uma rede de dispositivos, denominados nós sensores, que são distribuídos espacialmente e trabalham cooperativamente para comunicar informações coletadas em uma determinada área geográfica por meio de enlaces sem fio, tendo por objetivo relatar o monitoramento de fenômenos físicos, grandezas ou processos específicos [50].

Nesse tipo de rede, um nó sorvedouro é encarregado de receber os dados das leituras realizadas pelos nós sensores e repassar tais dados aos sistemas computacionais que realizam análise e armazenamento dos dados coletados. Tais sistemas computacionais dispõem de mecanismos de software que são capazes de efetuar o armazenamento dos dados recebidos, o gerenciamento e a configuração da RSSF e também é apto para tomar decisões em conformidade com a interpretação feita sobre esses dados [27]. Na Figura 2.1 é apresentada a estrutura básica de uma RSSF.

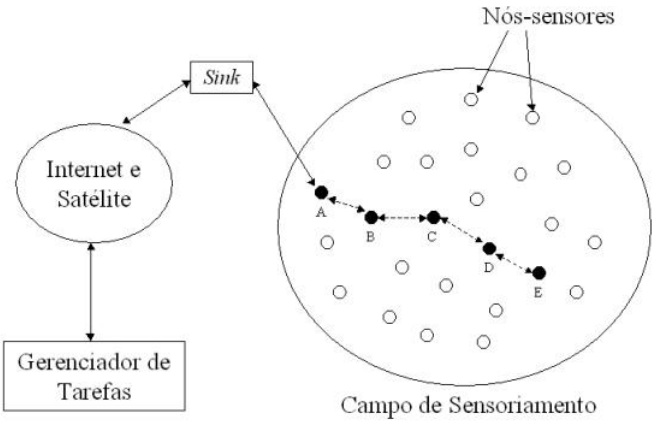

Figura 2.1: Estrutura básica de um Rede de Sensor Sem Fio [61]

Os nós sensores que compõem uma RSSF são dispositivos autônomos com capacidade de sensoriamento, processamento e comunicação [6]. Em geral, tais dispositivos têm por composição cinco elementos: um processador, que fornece funções de gerenciamento e processamento de dados, sensores capazes de detectar grandezas (ex: temperatura, umidade, luz etc), uma memória, que é utilizada para armazenar programas (instruções executadas pelo processador) e dados (medições dos sensores), bem como um transceptor integrado por um rádio sem fio, além de uma fonte de alimentação, em que comumente é utilizada uma bateria recarregável [50]. Devido ao fato de que os nós podem ser implantados em ambientes remotos e hostis, é comum as redes possuírem sensores que utilizem pouca energia e que empreguem mecanismos internos para prolongar a vida útil da rede [4]. Na Figura 2.2 são apresentados os componentes básicos de um nó sensor.

Como já tratado, as RSSFs possuem algumas vantagens quando comparada a soluções mais convencionais que utilizam, por exemplo, redes cabeadas, dentre elas pode-se destacar: menores custos, maior flexibilidade e maior facilidade de implantação. Dadas as evoluções tecnológicas e os avanços na elaboração dos sensores, que se tornam cada vez mais inteligentes, a implantação das RSSFs se torna cada vez mais recorrente [50]. De forma geral as RSSFs podem ser aplicadas em áreas de segurança, controle, manutenção de sistemas complexos e monitoramento de ambientes externos e internos [6]. Dentre estas aplicações temos em destaque:

- **Aplicações Militares**: Características de implantação rápida, auto-organização e to-

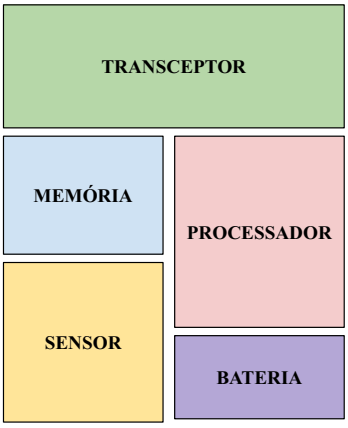

Figura 2.2: Estrutura básica de um sensor. Adaptado de [45]

.

lerância a falhas, fazem com que a RSSF seja uma tecnologia bastante promissora na área militar. Algumas das aplicações militares das RSSF são: monitoramento de equipamentos e munições ; vigilância de campo de batalha; reconhecimento de forças e terrenos inimigos; avaliação de dano de batalha; e detecção e reconhecimento de ataques nucleares, biológicos e químicos [43].

- **Aplicações Ambientais**: Aplicações para monitoramento de variáveis ambientais são bastante recorrentes em RSSFs. Uma dessas aplicações se dá na detecção de incêndios florestais. Nesses casos, os nós sensores podem ser estrategicamente implantados em uma floresta, e então transmitir a origem exata do fogo para os usuários finais, antes que o incêndio se alastre de maneira incontrolável [70]. Também em outros tipos de aplicações as RSSFs podem ser utilizadas, como: agricultura de precisão, pesquisa meteorológica ou geofísica, estudo da poluição, entre outros.

- **Aplicações Médicas**: Na área da saúde, as RSSFs também ganham força e espaço. Por exemplo, no monitoramento de dados fisiológicos humanos, em que informações fisiológicas são capturadas pelas redes sensores, podendo ser armazenadas por um longo intervalo de tempo e serem utilizadas para exploração médica [23]. Outros exemplos

incluem a administração de medicamentos em um hospital, rastreamento e acompanhamento de médicos e pacientes, diagnósticos, entre outros.

- **Aplicações Industriais**: Na área industrial as redes sensores sem fio destacam-se, tendo crescimento relacionado ao domínio do conceito de Indústria 4.0. Quando trata-se de aplicações, pode-se definir a taxonomia das RSSF industriais dividindo-as em três grandes grupos [15]:

  - **Sensoriamento ambiental:** Nesse primeiro grupo as aplicações exploram o monitoramento de parâmetros críticos para o funcionamento dos processos industriais, envolvendo monitoramento de risco, poluição e segurança dos processos, produtos e das pessoas envolvidas;
  - **Monitoramento de Condição:** Nessa classe a aplicação da rede pode se dar num monitoramento estrutural, onde sensores são instalados na infraestrutura da indústria, ou seja, em túneis, pontes, entre outros. Essa categoria também engloba monitoramento de equipamentos, com o intuito de detectar possíveis falhas de máquinas que provoquem prejuízos na produção;
  - **Automação de Processos:** Nesse último grande grupo, as aplicações da rede estão relacionadas com a avaliação da qualidade dos produtos que estão sendo produzidos no processo industrial, além de avaliação da utilização dos recursos industriais (como exemplo: água, energia). Além disso, as RSSF também podem ser aplicadas na melhoria de processos, utilizando atuadores e sensores para controlar e otimizar o processo de produção industrial.

A despeito de determinadas vantagens e um heterogêneo conjunto de aplicações, as redes sensores sem fio dispõem de desafios que vão além das redes convencionais. Limites de recursos como memória, energia, e capacidade computacional são obstáculos que afetam a rede de sensores [50]. Aliados a essas questões, problemas quanto a soluções de Controle de Acesso ao Meio (*Medium Access Control* - MAC) também ocorrem, como: colisões, sobrecarga de pacotes de controle e escutas ociosas [44].

Além desses aspectos, questões relacionadas à confiabilidade de qualidade de informações surgem de maneira desafiadora nas RSSFs. Problemas relacionados à segurança, como

interferências externas, falsificações de dados e monitoramento de terceiros, influenciam na precisão dos dados até a chegada ao usuário final [57]. Além disso, fatores relacionados a falhas de hardware, envolvendo calibração do sensor, defeitos no rádio transmissor, entre outros, interferem na qualidade de serviço da rede, necessitando de mecanismos que percebam e reajam a esse tipo de contrariedade [47].

Considerando tal conjuntura, as soluções propostas de qualquer nova abordagem ou protocolo, seja de camada física, de rede ou aplicação, devem estar atentas às variáveis, limitações e características mais relevantes concernentes às RSSFs, de maneira a atender aos serviços requisitados pelos usuários e lidar com as restrições impostas.

2.2 Qualidade de Serviço em Redes de Sensores Sem Fio

A qualidade de serviço (*Quality of Service* – QoS) é um termo técnico que pode possuir significados diferentes a depender da conjuntura e do contexto [42]. No entanto, de maneira geral, pode-de definir QoS como um conjunto de requisitos de serviço a serem cumpridos por uma rede durante o transporte de um fluxo de dados, ou seja, a capacidade da rede de garantir serviços prestados aos usuários de maneira satisfatória [39].

A QoS em redes de sensores sem fio é classificada sob duas perspectivas:

- **Específica de Aplicação**: Nessa perspectiva são considerados como parâmetros de QoS questões relacionadas à implantação, precisão de medição dos sensores, cobertura e números de sensores ativos. Requisitos que estão diretamente relacionados à qualidade das aplicações [42].

- **Específica de Rede**: Nessa perspectiva, o enfoque é dado na forma como a rede de comunicação de apoio consegue atender ao uso eficiente dos recursos de rede. Assim, a preocupação se dá com a forma em que os dados são entregues ao sorvedouro e seus requisitos correspondentes [42].

A utilidade primordial das RSSFs é a capacidade de gerar dados brutos, conseguir processar esses dados na rede, extrair informações pertinentes especificadas pela aplicação e entregar as informações geradas ao destino estabelecido. Percebe-se que, se estamos envolto

do conceito de QoS, a qualidade dos dados que trafegam pela rede gerando informação é de suma importância.

É importante pontuar que os conceitos de informação e dados são distintos apesar de estarem entrelaçados. Os dados referem-se a "partes" brutas monitoradas (leituras dos sensores), já a informação se caracteriza por um dado coletado, interpretado por processamento e requerido por uma determinada aplicação. Assim, sem uma coleta de dados precisa e um processamento adequado, as informações geradas pela rede serão afetadas e prejudicadas, aumentando a probabilidade de inconsistências.

Muito dos problemas que ocorrem entre os dados coletados dos sensores têm por origem de características de rede ou do *hardware* do nó. Tais erros apresentam-se de modos distintos, um deles se dá através de anomalias. Dado que as anormalidades podem possuir causas diversas, desde eventos legítimos do fenômeno que está sendo analisado, até falhas e ataques de terceiros, mecanismos que identifiquem e determinem a origem dos mesmos são desafios para fornecer uma maior QoS nas RSSF. Dessa forma, métodos que consigam detectar e categorizar as anomalias são valorosos para uma melhora na qualidade de informação das redes [37].

2.3 Detecção de Anomalias em Redes Sensores Sem Fio

Anomalias são pontos, ou um conjunto de pontos de dados, que não se adéquam, definidamente, ao comportamento padrão do grupo de dados. Na Figura 2.3 são ilustradas anomalias em um conjunto de dados bidimensional simples. Os dados têm duas regiões normais, N_1 e N_2, uma vez que a maioria das observações reside nessas duas regiões. Pontos que estão suficientemente distantes dessas regiões, por exemplo, os pontos O_1 e O_2, e pontos na região O_3, são considerados anomalias [11].

As anomalias podem surgir nos dados por diversas razões, como por exemplo, atividades maliciosas (invasão de sistemas computacionais), danos ao sistema, efeitos naturais de degradação, entre outros. Porém, todas essas razões possuem uma característica em comum, que é o interesse dos analistas, isso por que, as anomalias nos dados sempre irão traduzir algum tipo de informação relevante do que está ocorrendo na aplicação que está sendo executada. Por exemplo, um padrão de tráfego anômalo em uma rede de computadores pode

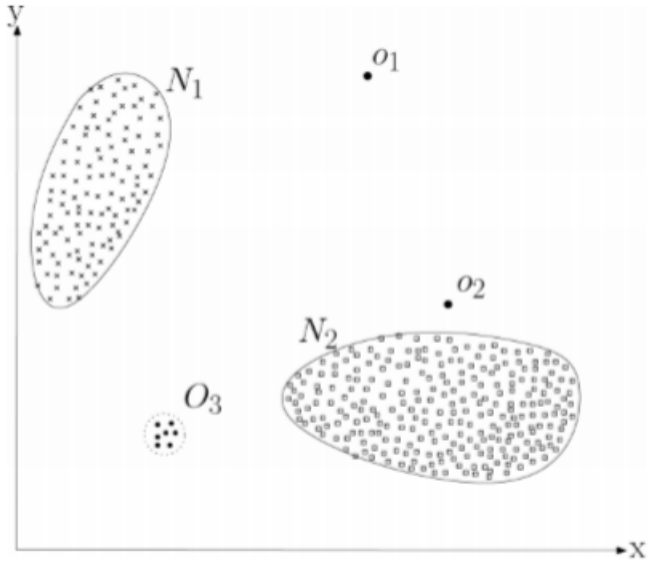

Figura 2.3: Exemplos de anomalias em um conjunto de dados bidimensional [11]

.

significar que um computador invadido está enviando dados confidenciais para um destino não autorizado [28]. Uma imagem anômala da ressonância magnética pode indicar a presença de tumores malignos [59]. Anomalias nos dados de transações com cartão de crédito podem indicar roubo de identidade ou cartão de crédito [2]

Determinados desafios ainda são recorrentes na detecção de anomalias, o que a torna em certo nível complexa, como por exemplo a definição de uma região que determine com precisão o que é um comportamento normal, visto que o limite que separa o comportamento padrão e o anormal em geral é tênue. Outra questão se dá nos ataques maliciosos, que cada vez mais se adaptam comportamentalmente, fazendo com que atividades anômalas sejam similares às atividades normais. Assim como dados que contenham ruídos e que, devido a isso, se assemelhem a anomalias reais que indicam eventos, dificultando a distinção e remoção das anormalidades

2.3.1 Características da Detecção de Anomalias

Uma das características mais fundamentais de qualquer técnica de detecção de anomalias é a natureza dos dados de entrada. A entrada é geralmente uma coleção de dados que pode ser descrita através de um conjunto de atributos. Cada conjunto de dados pode ser classifi-

cado como univariado, em que a instância consiste em apenas um atributo, ou multivariado, sendo a instância composta de vários atributos. Esses atributos podem ser classificados em diferentes tipos, como binário, categórico ou contínuo. Em casos de conjuntos de dados multivariados, todos os atributos podem ser de um tipo ou podem ser uma mescla de diferentes tipos de dados [11].

A natureza dos atributos a serem analisadas determina a aplicabilidade da técnica a ser utilizada na detecção de anomalias. Além disso, os dados de entrada também podem ser categorizados quanto à relação existente entre os conjuntos de dados. Na maioria dos casos, as técnicas de detecção lidam com dados pontuais, em que não se assume relação entre conjuntos. Outro aspecto importante, no que tange à detecção de anomalias, é a natureza das anomalias. A Figura 2.4 exemplifica os tipos existentes de anomalias, em que uma anormalidade pode ser classificada em três tipos básicos[11]:

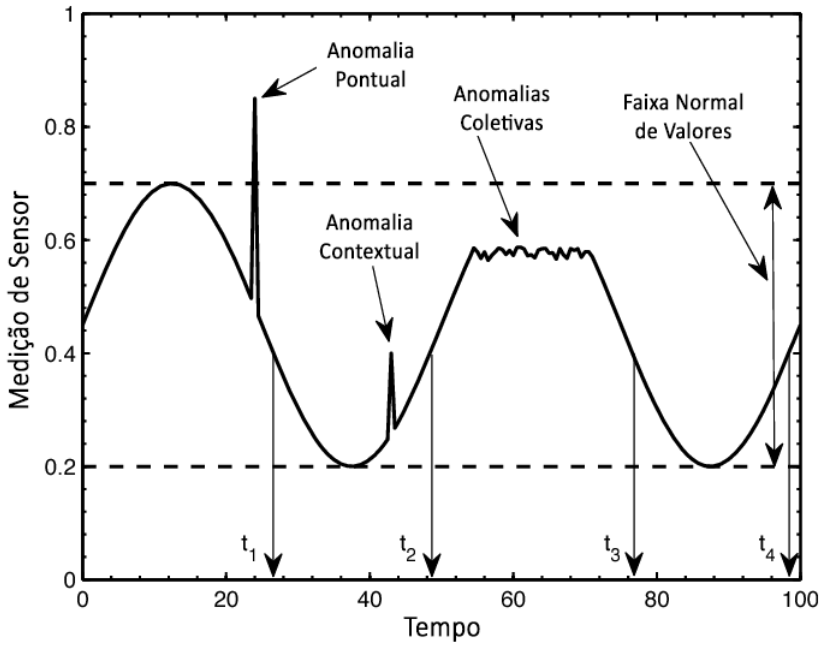

Figura 2.4: Exemplos de tipos de anomalia. Adaptado de [41]

- **Pontuais:** Trata-se de quando os pontos de dados individuais se desviam do restante do conjuntos de dados. É considerado o tipo mais simples e mais abordado pelos algoritmos existentes;
- **Contextuais:** É quando um determinado dado é anômalo em um determinado con-

texto, mas não em outro, sendo o contexto determinado pela estrutura do conjunto de dados;

- **Coletivo:** Ocorre quando um grupo de dados relacionados é anômalo em comparação a todo o conjunto de dados. As instâncias individuais de dados podem não ser consideradas anomalias, porém, quando ocorrem de forma coletiva, são classificadas como anômalas.

As técnicas de detecção de anomalias também são classificados, e podem ser categorizadas quanto ao tipo de dados de treinamentos esperados [11]:

- **Supervisionada:** Nesse tipo de técnica um grupo de dados é coletado compondo um conjunto para treinamento, de forma que posteriormente tais dados sejam rotulados em classes específicas. Em seguida, cria-se um modelo que aprende a classificar dados de acordo com os elementos que foram utilizados na fase de treinamento;

- **Semi-supervisionada:** Esse tipo de técnica se assemelha à anterior, pois também requer uma fase de treinamento. Nesse modo, o conjunto de treinamento contém apenas registros de dados considerados normais. Assim, um dado é rotulado como anômalo se o mesmo se distanciar do modelo normal treinado. Como nesse caso não se exige rótulos para a classe de anomalia, tais técnicas são mais aplicáveis do que as supervisionadas;

- **Não-supervisionada:** Nesse tipo de técnica, os conhecimentos a *priori* gerados em um grupo de treinamento não existem. Assim, o objetivo dessa categoria é descobrir um padrão implícito em um conjunto de dados não rotulados.

A maneira como as anomalias são apontadas também é uma perspectiva importante na detecção de anormalidades, em que duas saídas são possíveis para os algoritmos de detecção: pontuações e rótulos. Quando se trata de pontuações, como o próprio nome sugere, os algoritmos atribuem pontuações de anomalia a um determinado item dos dados e dependendo do grau atribuído à instância, o mesmo pode ser considerado uma anomalia. Já os rótulos apenas determinam se um dado é anômalo ou não (*True/False*) [11].

Além de todas essas categorizações, as técnicas podem ser aplicadas sobre dois modos: *online* e *offline*. No modo *online*, novos dados são continuamente introduzidos e as anomalias

devem ser detectadas em tempo real. Já o modo *offline*, refere-se a quando as anomalias são detectadas em um conjunto de dados já existentes. No contexto das RSSFs, a aplicação das técnicas em ambiente real ocorre no modo *online*, necessitando de mecanismos que consigam avaliar constantemente os dados recebidos e aplicar as técnicas de detecção através de janelamentos [11] [3].

2.3.2 Anomalias em Redes de Sensores Sem Fio

Os dados coletados a partir das RSSFs podem ser acometidos por anomalias. Tais anormalidades possuem origens diversas, como falhas de *hardware* dos nós sensores, problemas com *softwares* e de transmissão, fatores dinâmicos ambientais, intrusões, etc. Desta forma, a volatilidade das RSSFs antagoniza com os objetivos das aplicações mais ambiciosas do paradigma, e a confiabilidade da entrega dos dados se torna um fator de grande desafio [27] [30].

Dentre esses desafios, temos problemas relacionados à grande quantidade de dados coletados dos nós sensores, o que dificulta a identificação de anormalidades nas medições realizadas. Além disso, as técnicas de detecção necessitam atuar no modo *online*, de forma a detectar automaticamente as falhas e informar ao administrador da rede, para que o mesmo tome as medidas oportunas. Outro aspecto é que as técnicas de detecção devem ser modestas no que se refere à demanda por poder computacional, devido às limitações de recursos dos nós sensores [60].

No geral, o diagnóstico de anomalias ocorre a partir de ferramentas convencionais de gerenciamento de rede. Inicialmente, a principal motivação para este diagnóstico é mapear os desvios ocorridos, para então a partir disso poder investigar as possíveis causas, e assim poder propor ações corretivas. Com base nesse processo, as anomalias nas RSSFs são divididas em três tipos de categorias [24]:

- **Anomalia de Rede:** São problemas relacionados à comunicação que surge entre os nós da rede. Os sinais típicos em geral são um aumento ou diminuição inesperada na quantidade de pacotes que transitam na rede [24].
  - **Perda de Conectividade:** Considerado o tipo mais simples de anomalia de rede, a perda de conectividade é caracterizada pela interrupção de pacotes de entrada

entre dois ou mais nós, ou seja, é a falta de recepção de pacotes de um nó ou de um grupo de nós (podendo ser até todos os nós) por um determinado intervalo;

- **Conectividade Intermitente:** Nesse caso, as anomalias ocorrem devido à alta variação da frequência de recepção de dados de certos nós, em referência a um limite de estabilidade de enlace determinado pelo operador da rede. Esse tipo de falha também pode abranger todos os nós da rede;
- ***Loops* de Roteamento:** Os *loops* de roteamento ocorrem quando pacotes são retransmitidos por determinados nós na rede e retornam ao nó de origem. A detecção desse tipo de anomalia é considerada complexa, em que os protocolos para solucionar tais problemas devem envolver bastante sobrecarga de comunicação, afinal vários nós estão incluídos no processo;
- **Tempestades de *Broadcast*:** Nesse último caso de anomalias de rede, muitos ou todos os nós podem ser afetados. O que ocorre é que determinados nós que perderam a sua conexão com seus pares de roteamento podem decidir transmitir seus pacotes de maneira contínua, em busca de caminhos alternativos para a estação base, gerando uma sobrecarga na rede.

- **Anomalias do Nó:** As anomalias em nível de nó estão relacionadas a falhas no *hardware* ou *software* de um determinado nó, não tratando de questões de comunicação entre nós vizinhos [24].
  - **Problemas de Bateria:** Esse tipo de anomalia é caracterizado pela falta de fornecimento de energia, ocasionado pela falha ou esgotamento da bateria do nó. Tais problemas na bateria existem por duas possibilidades: carga insuficiente da bateria ou falha do *hardware* da bateria;
  - **Falha do Nó:** Nesse caso, as anomalias surgem de problemas relacionados à memória, à *CPU* ou ao rádio do nó, que podem entrar em um estado de bloqueio. Esse tipo de situação pode ocorrer por causa de componentes de *hardware* defeituosos ou de uma integração falha entre *softwares* e os componentes.
- **Anomalias dos Dados:** As anomalias dos dados estão ligadas a irregularidades estatísticas nos dados sensoreados. Tais irregularidades podem ser causadas por falhas

de segurança, como intrusão de rede, bem como por variações ambientais. Além de poderem ser causadas pelo *hardware* defeituoso do sensor do nó. Vale ressaltar que problemas do sensor são considerados anomalias de dados em vez de anomalias de nó, isso porque as anormalidades se apresentam em valores de dados errados e não em uma falha ou desempenho reduzido do nó [24].

- **Anomalias Temporais:** Esse tipo de anomalia é caracterizada por um dos seguintes fatores: alta variabilidade nas leituras dos sensores ou a falta de mudança na leitura dos sensores. A alta variabilidade das leituras dos nós podem significar grandes mudanças no ambiente detectado ou problemas no sensor. Já amostras do sensor que permanecem as mesmas em um intervalo, podem indicar um estado de bloqueio ou que o equipamento não esteja conseguindo obter novas amostras;
- **Anomalias Espaciais:** As anomalias de dados espaciais são irregularidades que podem ser detectadas comparando os valores de um determinado nó sensor com os nós sensores adjacentes. Se a medição de uma grandeza, como ar, temperatura ou umidade, feita por um nó, difere substancialmente dos nós vizinhos, significa que, provavelmente, os dados sejam espacialmente anômalos;
- **Anomalias Espaço-Temporais:** O último tipo de anormalidade de dados são as anomalias espaço-temporais, em que se combina variações espaciais e temporais, envolvendo mais de um nó, requerendo uma interação entre os nós, a fim de estabelecer a existência da anomalia.

Dadas as características das anomalias existentes nas RSSFs, três segmentos de detecção são estabelecidos na literatura: Detecção de Eventos, que está relacionada a alterações nas leituras causados por mudanças físicas do estado no ambiente sensoreado, Detecção de Falhas, que são produzidas por nós defeituosos, e a Detecção de Intrusão, que são as ocasionadas por invasores. A Figura 2.5 apresenta a divisão dos contextos de detecção de anomalias em Redes de Sensores Sem Fio. [5]

Além desses fatores, as técnicas de detecção aplicadas às RSSFs também podem ser classificadas de acordo com sua arquitetura, para esse contexto são definidos três grupos:

- **Centralizada**: Nesse tipo de arquitetura, a detecção ocorre na estação base (ou sorvedouro). Os sensores são encarregados de fazer as leituras e enviarem a informação até

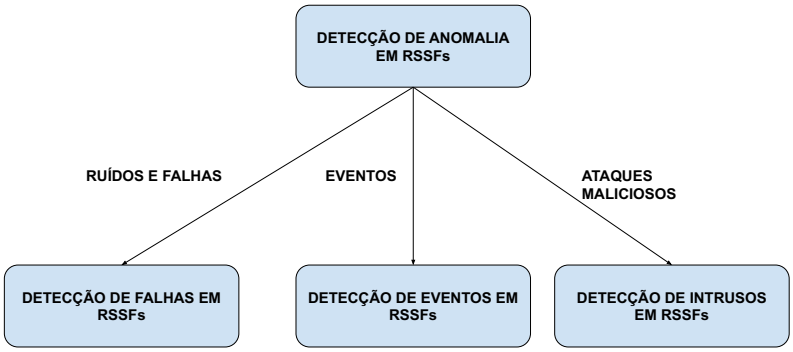

Figura 2.5: Tipos de detecção de anomalias em Redes de Sensores Sem Fio. Adaptado de [5].

o sorvedouro, que por possuir maior poder computacional pode processar as informações e detectar as anomalias com maior precisão. Por outro lado, tal arquitetura possui como gargalo uma sobrecarga no tráfego e de consumo de energia;

- **Distribuída**: Como o próprio nome sugere, uma abordagem distribuída compartilha a execução. Esse processamento pode ocorrer nos nós locais e estação base, ou também componentes intermediários podem ser criados, componentes estes chamados de *Cluster Heads*, que centralizam a informação de um determinado subgrupo de sensores da região monitorada;

- **Local**: Já a arquitetura local se refere ao conjunto de técnicas que processam a detecção nos nós sensores finais, que neste caso não apenas coletam a informação, mas processam e identificam as anomalias. Neste tipo de cenário, os mecanismos e protocolos criados têm como desafio o gargalo computacional, devido ao baixo poder de processamento inerente aos nós sensores.

Quando se trata de análise dos dados e de informação, compreender a origem da anomalia é essencial, uma vez que por meio da identificação da causa das anormalidades, a construção de mecanismos para solucionar determinados problemas na rede se torna mais fácil. Conseguir definir se o que ocorre naquela região é um evento ou alguma falha sistêmica está ocorrendo, é de suma relevância para a confiabilidade e otimização da rede.

Por exemplo, se uma rede de sensores sem fio é distribuída em uma floresta, com o intuito de detectar possíveis incêndios sensoreando a temperatura do ambiente, e acaba recebendo um fluxo de dados considerados anômalos, é de máxima importância que o nó sorvedouro e o operador de rede consiga analisar qual a probabilidade desses dados extremos estarem vindo de uma falha na medição do sensor ou que a origem realmente seja de um incêndio florestal que esteja ocorrendo na região. De forma que discernindo o motivo da aparição desses dados discrepantes, soluções mais precisas e coerentes possam ser aplicadas, evitando assim perdas de recursos e tempo.

2.4 Lógica Fuzzy

A lógica fuzzy (ou difusa) consiste em uma forma de lógica multivalorada, em que os valores verdade das variáveis podem ser qualquer número real entre 0 (falso) e 1 (verdadeiro). Ela se distingue da lógica booleana, onde os valores lógicos estão restritos a serem 0 ou 1. Assim como é o raciocínio humano, a lógica difusa consiste em modelar um problema de modo aproximado em vez de preciso [31].

Na lógica fuzzy é possível trabalhar com informações incertas, que comumente são utilizadas no dia a dia (como responder questionamentos com talvez, mais ou menos, às vezes, depende, etc.). A Figura 2.6 apresenta um exemplo sobre a diferença entre o conceito de resposta para a lógica booleana e para a lógica difusa, considerando a utilidade de um artigo lido [67].

Figura 2.6: Exemplos de respostas consideradas para lógica fuzzy [67]

.

Derivado desse conceito, a construção de um sistema de lógica de fuzzy depende de cinco componentes: Entrada e Saída Crisp, Fuzzificação, Regras Fuzzy, Inferencia Fuzzy, Defuzzificação. A Figura 2.7 ilustra o funcionamento de um sistema fuzzy.

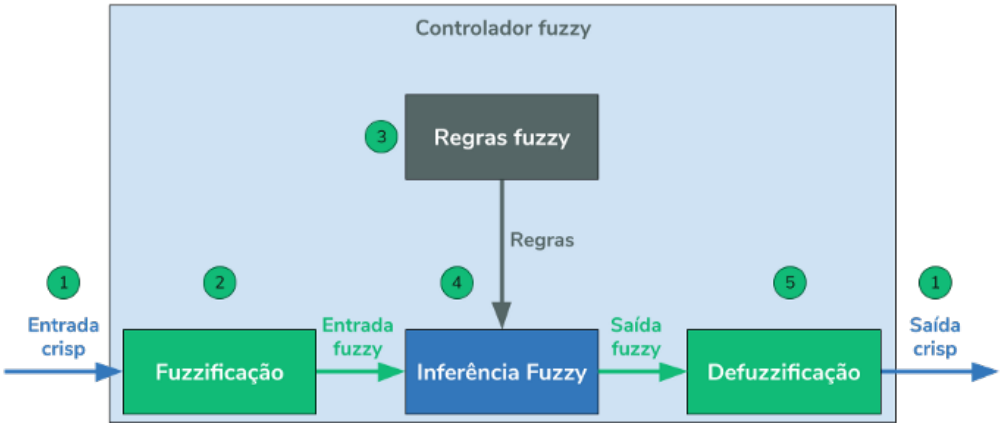

Figura 2.7: Ilustração de um Sistema Fuzzy [67].

2.4.1 Entrada Crisp

Toda variável fuzzy detém um valor chamado *crisp*, que consiste em um número dentro de um domínio que foi pré-estabelecido. Esse domínio é chamado de Universo, e determina uma faixa de valores em que a variável se encontra, sendo necessária a definição de limites mínimos e máximos. Assim, as entradas para o sistema de controle fuzzy serão os valores crisp de cada variável [67].

2.4.2 Fuzzificação

O primeiro momento do sistema fuzzy ao receber os valores de entrada crisp, é a fuzzificação, que consiste em transformar o valor crisp em um valor fuzzy. Para isso, dois conceitos são importantes: o termo e a função de pertinência [67].

- **Termo:** O termo são as formas que podem ser utilizadas para descrever uma variável fuzzy. Todos esses termos são utilizados para definir qual será o valor da variável fuzzy. Como por exemplo: forte, fraco, mediano, ruim, muito bom.
- **Função de pertinência:** Todo termo possuirá uma função de pertinência, que tem como objetivo definir como a entrada crisp será mapeada dentro de uma escala (em

uma faixa que vai de 0 a 1), que aponta para o grau de pertinência do valor para o termo em questão. Ou seja, a função define o pertencimento de um elemento a um determinado conjunto. A Figura 2.8 exemplifica algumas funções de pertinência.

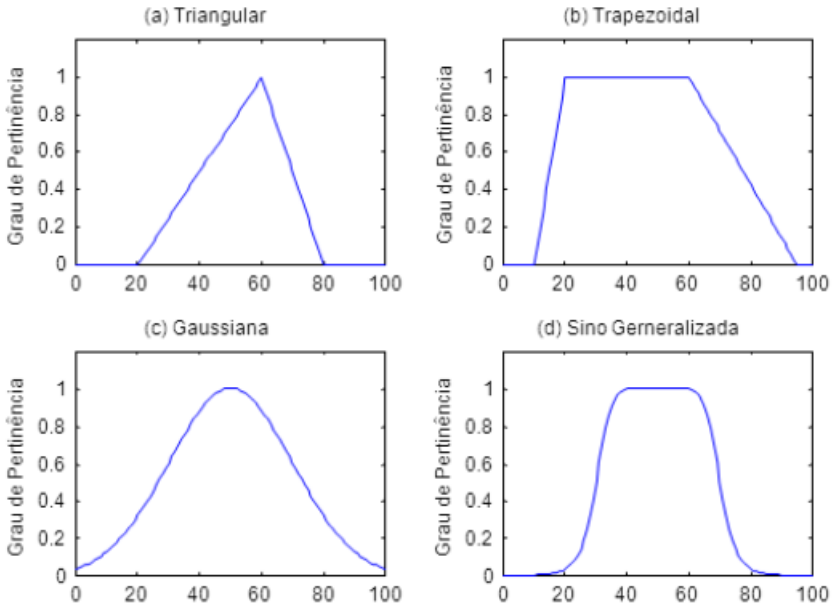

Figura 2.8: Exemplos de funções de pertinência [11]

.

2.4.3 Inferência Fuzzy

Após a fuzzificação, os valores fuzzificados são aplicados a um conjunto de regras de inferência que estabelecem relações de causa e efeito entre as variáveis de entrada e de saída. As regras de inferência que melhor representam os valores de entrada fuzzificados serão então ativadas, determinando assim um conjunto de hipóteses sobre o comportamento do problema modelado. As regras fuzzy são utilizadas com o objetivo de interligar diferentes variáveis fuzzy, de maneira a descrever como tais variáveis estão relacionadas entre si. Tais regras são expressas por meio de uma declaração SE/ENTÃO. A declaração SE é chamada de **antecedente** e a ENTÃO é chamada de **consequente** [31][67]. Em que os elementos de fuzzificação (termos e funções de pertinência) devem ser estabelecidos para cada declaração. O padrão para a construção de uma regra segue então a seguinte forma:

SE a está A_i **ENTÃO** b está B_i,

Os conjuntos fuzzy resultantes do processamento nas regras ativadas são agregados em um único conjunto para gerar o valor de saída do sistema. O método de inferência fuzzy mais comumente utilizado é o chamado método Mamdani. O primeiro passo na inferência fuzzy utilizando o método Mamdami é fornecer os graus de pertinência de cada variável (de acordo com os conjuntos fuzzy) para as regras que compõem o sistema e, assim, encontrar a região resultante. Após isso, agregam-se as saídas das regras para o processo de defuzzificação [32].

2.4.4 Defuzzificação

A etapa de defuzzificação tem por objetivo converter a saída da inferência fuzzy, que foi gerada baseada nas regras estabelecidas e nas funções de pertinência, em um valor crisp. O método mais utilizado para tal é o centróide, que pode ser aplicado para valores contínuos ou discretos. Para valores discretos, calcula-se a média ponderada de acordo com o grau de pertinência para para a distribuição de possibilidades de saída do modelo. Já para valores contínuos, torna-se necessária a aplicação de integrais [67][32], como demonstrado nas equações a seguir.

$$\mathrm{X} = \frac{\int_a^b \mu_A(x) x \, dx}{\int_a^b \mu_A(x) \, dx} \tag{2.1}$$

$$\mathrm{X} = \frac{\sum_{x=a}^{b} \mu_A(x) x}{\sum_{x=a}^{b} \mu_A(x)} \tag{2.2}$$

em que X é o valor crisp que sairá do sistema, x é o exemplo observado no momento e $\mu_A(x)$ é o grau de pertinência observado.

2.5 Considerações Finais do Capítulo

Neste capítulo, foram apresentados os principais conceitos relacionados às redes de sensores sem fio, passando pelos conceitos de qualidade de serviço e detecção de anomalias em redes de sensores sem fio, assim como o conceito de lógica fuzzy, que é um fator essencial para

a construção da proposta desta dissertação. No capítulo seguinte, serão apresentados os trabalhos relacionados à categorização de anomalias em redes de sensores sem fio, com destaque para suas vantagens e limitações.

Capítulo 3

Trabalhos Relacionados

Neste capítulo, são discutidos os trabalhos que abordam o problema detecção de falha e detecção de eventos em Redes de Sensores sem Fio (RSSF). Os trabalhos selecionados foram encontrados a partir da realização de consultas aos sites de busca Google Scholar, IEEE Xplore Digital Library e ACM Digital Library, consistindo de trabalhos publicados desde 2004. As principais palavras-chave consideradas foram: "data", "fault", "wsn", "event", "detection", "anomaly". As palavras-chave foram verificadas tanto no título dos trabalhos como nos conteúdos.

O desenvolvimento de técnicas e algoritmos para a detecção de anomalias em RSSF possui vasto estudo no âmbito acadêmico. Todavia, se tratando da categorização das anomalias, conseguindo apontar a ocorrência de falhas ou eventos, a restrição de estudos é maior. Além de que, mecanismos existentes possuem limitações para definir com confiança em que categoria as anomalias existentes se enquadram.

3.1 Detecção de Eventos

A detecção de eventos está relacionada à capacidade de identificar se algum acontecimento em relação à variável monitorada está ocorrendo de forma a gerar valores fora do esperado. Nas técnicas mais comuns, os eventos são detectados usando um limite definido pelo usuário [63][64]. Nestes mecanismos, um alarme de evento é ativado quando as leituras dos sensores são maiores ou menores dos limiares pré-estabelecidos. Por exemplo, o trabalho descrito em [18], através de limiares estabelecidos a partir de leituras de sensores

de movimento e sensores acústicos consegue rastrear veículos em uma determinada região a partir de alterações dos valores dos dados observados.

Seguindo ainda tal contexto, o trabalho descrito em [9] propõe um modelo online distribuído para detecção de eventos em RSSFs utilizando sensoriamento comprimido e iterações, de forma a capturar o estado atual do ambiente ou do fenômeno monitorado. O algoritmo impõe pesos às leituras dos sensores, e baseado em um limiar conclui se existe ali um evento não. Para superar limitações de faixa de limites pré-definidos, propostas como o de [34] surgem de forma que o valor do limiar é encontrado dinamicamente usando um método de janela deslizante, em que os valores mais recentes definem o limiar a ser estabelecido, aliado a um sistema fuzzy que consegue estabelecer os melhores padrões para definir a existência de um evento.

Ainda assim, a definição de limiares, mesmo que dinamicamente, pode ser limitante, e para não ficar preso a essa restrição, trabalhos como de [33] funcionam agregando dados dos nós sensores na estação base em um mapa de dados e detecta os eventos através da combinação de uma série temporal de valores do mapa de dados com as propriedades conhecidas de certos eventos.

Por consequente, modelos de aprendizado de máquina para detecção de eventos também são empregados, utilizando os próprios dados ou informações anteriores coletadas. Em geral, esses modelos são aplicados em arquiteturas distribuídas. São os casos de [58] e [7]. O trabalho descrito em [58] propõe uma abordagem de aprendizado de maquina distribuído por *ensemble*, em que todos os *Cluster Heads* são utilizados para armazenar e guardar informações do *cluster*. A estação base recebe todas as informações e o processamento das informações é realizado, através de uma verificação dos dados com o registro estatístico por uma máquina de vetor de suporte. Já o trabalho descrito em [7] usa uma abordagem baseada na detecção de eventos usando classificadores de árvore de decisão, que executa em nós sensores individuais, aplicando uma votação para chegar a um consenso entre as detecções feitas por todos os nós sensores.

Por fim, tem-se o trabalho descrito em [56], que trata de uma implementação de uma técnica baseada no SVM, mais especificamente uma expansão da técnica, denominada QS-SVM, para detectar dados anormais. Em seguida, é utilizado o conceito de Sensgru, que consiste no agrupamento de medidas de diversos sensores em um único nó, para assim esta-

belecer se existe um evento naquela localidade.

Em linhas gerais, diversos mecanismos com diversos conceitos são apresentados, porém quando tratamos de uma abordagem distribuída/hibrida se percebe o estabelecimento do conceito de agrupamento e/ou similaridade, em que dada uma região, o que determinará a existência de um evento ou não é a alteração nos padrões dos dados de todos os nós da região. Conceitos que apenas consideram as alterações nos dados nos sensores individualmente, possuem o risco de apontar a existência de eventos em momentos em que os sensores apenas estão com problemas de *hardware* ou *software*.

3.2 Detecção de Falhas

Quando tratamos de detecção de falhas ou tolerância a falhas em RSSFs, é necessário compreender dois níveis. O primeiro nível está ligado à detecção de falhas utilizando conceitos de camadas inferiores. Em [55], informações do transceptor, do controlador e da bateria dos sensores de uma vizinhança são utilizadas para determinar sensores falhos e em que localidade do *hardware* essa falha está concentrada. Assim como em [10], que propõe um esquema de classificação e gerenciamento de falhas baseado em um sistema *fuzzy*, utilizando informações do hardware, e assim categorizando os sensores aos níveis de falhas existentes.

No caso de tratarmos de detecção de falhas em camadas de mais alto nível (tratando dos dados), podemos classifica-los em algumas categorias. Em [26] temos abordagens centralizadas, em que o nó sorvedouro assume a responsabilidade pelo gerenciamento de falhas de toda a rede, agregando as informações dos sensores e tomando as decisões. Por outro lado temos os mecanismos distribuídos, [12], [22], [69], que utilizam conceitos de detecção espacial, aliado a modelos bayseanos de probabilidade determinando quais sensores são falhos ou não. Em comum, esses trabalhos apesar de analisarem o grau de incidência de nós falhos, não propõem e nem conseguem averiguar meios de detecção de falhas em casos em que em uma determinada região ou cluster os sensores falhos se constituem maioria. Além disso, os trabalhos não visam averiguar a relação das técnicas com perda de dados e as variáveis existentes de um ambiente online, fatores preponderantes nos mecanismos que funcionam para o contexto das RSSFs.

3.3 Categorização de Anomalias

No que se trata de categorização de anomalias, o trabalho descrito em [16] apresenta um algoritmo baseado em clusterização aliado ao vizinho mais próximo, no qual os dados são agrupados em clusters, passando por uma detecção de anomalia, em seguida através da correlação espacial dos vizinhos mais próximos defini-se a existência de um erro do nó sensor ou de um evento, tendo como última etapa aplicar uma equação simples de quantidade de valores falhos em um determinado período, que determina a confiabilidade do sensor.

Ainda considerando o problema proposto, o trabalho de dissertação descrito em [5] define uma abordagem considerando RSSFs de larga escala, destacando-se por utilizar técnicas estatísticas por seu baixo custo computacional, além de agrupar tanto a identificação quanto a detecção de anomalias na mesma abordagem, diferente da maioria dos trabalhos relacionados ao problema, além de considerar a escalabilidade da rede, visto que a aplicação se dá para um ambiente de grande quantidade de nós. Porém não considera questões relacionadas a perda de dados e a incidência de nós falhos nas regiões monitoradas. Vale ressaltar também, que apesar do mecanismo considerar a existência de eventos, o enfoque do estudo em si está focalizado na identificação de nós falhos.

3.4 Considerações Finais do Capítulo

No melhor do nosso conhecimento, considerando o estado da arte atual, os trabalhos existentes para categorização de anomalias em RSSFs possuem quantidade limitada, além disso questões envolvendo ambientes online, relação com perda de pacotes e quantidade de sensores falhos são tratados de forma superficial ou não são investigados. Um dos pontos mais importantes é que para as abordagens de categorização e até detecção de sensores falhos apresentados no estado da arte, não são considerados contextos em que os sensores falhos sejam metade ou a maioria do *cluster*/vizinhança. Seguindo estas limitações, o presente trabalho, desenvolve uma nova abordagem para categorização de anomalias online em RSSFs, para identificar eventos e sensores falhos, considerando em certo nível contextos em que os sensores falhos são metade ou maioria da vizinhança, além de considerar tipos de anomalias distintas como pontuais e coletivas.

Capítulo 4

Proposta

O presente trabalho tem por objetivo desenvolver uma abordagem distribuída de detecção e categorização de anomalias *online* para ambientes de Redes Sensores sem Fio (RSSF). Dadas as limitações das soluções do estado da arte para esse tipo de problema, como discutido no capítulo anterior, a abordagem proposta neste trabalho procura agregar dois níveis de atuação: a primeira visa detectar anomalias, ou seja, encontra dados que fogem do padrão esperado e a segunda visa categorizar o tipo de anomalia, se a mesma se refere a uma falha de sensor ou um evento que ocorre na região monitorada. E assim, permitir a discussão de quais seriam os melhores caminhos para atuar como resposta ao tipo de anomalia encontrada no monitoramento.

A proposta tem como característica arquitetural uma composição hibrida. Neste sentido, a primeira etapa de detecção e categorização ocorre de forma interligada entre os sensores e o *Cluster Head*. Na esfera da detecção de anomalias, a abordagem é construída a partir de técnicas estatísticas ou de agrupamento, que encontram padrões nas informações recebidas em janelas de dados de tamanho pré-estabelecido, de modo a determinar a existência ou não de elementos considerados anormais em relação aos dados sensoreados pelo nó sensor que está sendo analisado. Para a categorização das anomalias, o modelo construído visa atribuir níveis de confiabilidade aos sensores, direcionando assim a determinação de que tipo de categoria as anomalias que surgiram pertencem. Assim, quanto menos confiável é um sensor, maior é a chance de termos uma anomalia/falha de nó.

A Figura 4.1 ilustra a arquitetura da rede de sensor sem fio considerada na proposta. Onde os elementos em vermelho representam os *Cluster Heads* (CH), os azuis os sensores e

em amarelo se é representado o nó sorvedouro ou estação base.

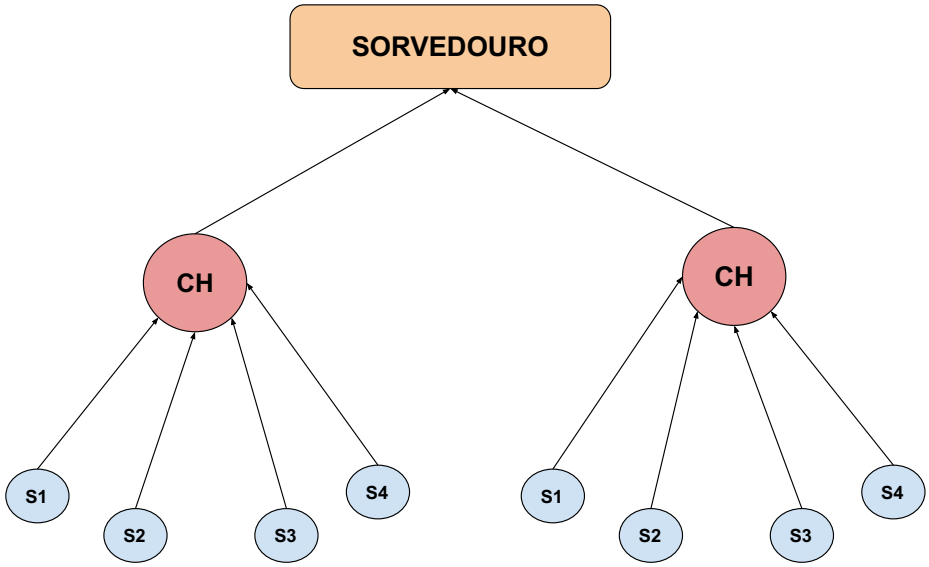

Figura 4.1: Arquitetura de Rede

.

4.1 Detecção e Categorização - Primeira Camada

Como já relatado, a primeira camada consiste em dois componentes essenciais: os **sensores** e o ***Cluster Head***. Os sensores possuem como função monitorar o ambiente de uma região (vizinhança), capturar as informações da variável estabelecida e enviar essas informações ao *Cluster Head*. O *Cluster Head*, como um elemento intermediário entre os sensores e a estação base (ou sorvedouro), possui algumas funcionalidades: ele executa o roteamento dos dados recebidos dos sensores para o sorvedouro, armazena a informação periodicamente por meio de janelas para a detecção de anomalias em cada sensor, além de exercer o papel de comparador de dados dos sensores e assim conseguindo, já nesta etapa, estabelecer algumas definições, como se há ou não um evento ocorrendo naquela vizinhança.

O funcionamento detalhado da primeira etapa, apresentado no diagrama da Figura 4.2, é estabelecido nos seguintes passos:

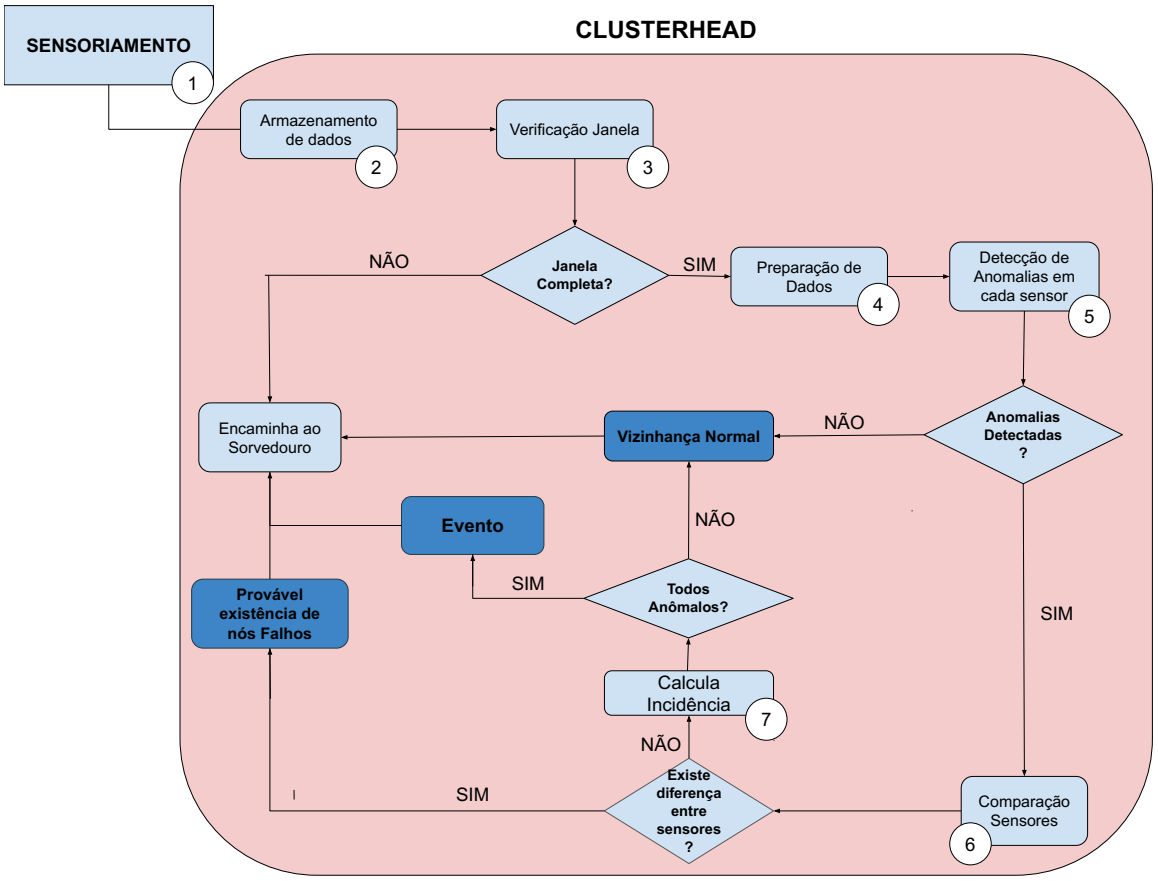

Figura 4.2: Diagrama da abordagem de categorização primeira etapa

.

1. Inicialmente os dados são sensoreados e enviados por meio de pacotes, que possuem a informação da variável além do identificador sequencial (*id*) do pacote, ao *Cluster Head*. Cada sensor envia seu pacote diretamente para o *Cluster Head*;

2. Após receber o pacote, o *Cluster Head* armazena o dado proveniente do nó sensor em seu respectivo *buffer*;

3. Através da informação do *id* do pacote, o *Cluster Head* define o que deve ser feito com aquele dado, verificando se a janela está completa ou não.

   (a) Caso o *id* do pacote de um sensor não seja equivalente ao tamanho da janela de

análise de dados estabelecida (*buffer*), o *cluster* mantém a informação armazenada e apenas faz o roteamento ao sorvedouro.

(b) Caso o *id* de um sensor seja equivalente ao da janela, o *Cluster Head* chama o processo de preparação dos dados.

4. No processo de preparação de dados, os pacotes perdidos e que não estão na janela dos sensores, são substituídos pela mediana dos valores, permitindo a detecção e a posterior comparação com os demais sensores. Os dados das janelas só são enviados ao processo de detecção no momento em que todos os outros sensores da vizinhança tenham enviado os seus pacotes referentes ao mesmo *id* e passado pelo processo de preparação de dados. Isso ocorre, pois o *Cluster Head* faz o processo de detecção nos sensores apenas após todos os nós da vizinhança terem enviado o pacote com id referente ao tamanho da janela estabelecida. Permitindo assim, a comparação entre sensores nos passos futuros.

5. Após isso uma técnica de detecção de anomalias é aplicada em cada janela de cada sensor, identificando assim a existência de anormalidades naquele período ou não. Aplicada a técnica de detecção se é realizada a análise da existência ou não de anomalias.

   (a) Caso nenhuma anomalia seja detectada nos sensores, a vizinhança é declarada normal e o status é enviado ao nó sorvedouro.

   (b) Caso alguma anomalia seja encontrada em algum sensor, o processo de comparação entre sensores é chamada.

6. Identificada a existência de anomalias em algum sensor, o processo de comparação é iniciado, no qual o conjunto de dados dos sensores são comparados, afim de determinar a similaridade de cada sensor com os demais.

   (a) Caso o conjunto de dados obtidos de algum sensor apresentar comportamento diferente dos demais, então existe uma maior probabilidade de que algum desses sensores estejam enviando informações errôneas. Portanto, um alerta de possível nó falho é enviado ao sorvedouro.

(b) Caso não exista diferença entre os sensores, então o nível de incidência de anomalias é analisado.

7. Tendo a informação dos sensores e de suas respectivas anomalias, incidência de sensores anômalos é analisada.

   (a) Caso todos os sensores possuam dados anômalos, então um provável evento está ocorrendo, e o alerta de evento é gerado e encaminhado ao sorvedouro.

   (b) Caso nem todos os sensores possuam anomalias, é provável que a técnica tenha falhado na detecção, gerando falsos alarmes, e desta forma se é considerada uma vizinhança normal.

Ao fim de todo o processo, os *buffers* que armazenam os dados no *Cluster Head* são esvaziados.

4.2 Componentes da Abordagem - Primeira Camada

A abordagem é composta por componentes genéricos que podem ser municiados de acordo com a preferência/conhecimento do agente que irá utilizar o mecanismo. Para a primeira etapa temos dentre esses componentes dois pontos que são pertinentes: A técnica de detecção, para identificar as anomalias e o mecanismo de comparação de dados dos diferentes nós sensores, para determinar o nível de similaridade entras as medições realizadas pelos diferentes nós sensores.

Nas sub-seções a seguir são descritas as implementações específicas de cada componente, que foram realizadas neste trabalho para avaliar a abordagem proposta.

4.2.1 Técnica de Detecção

Nesse trabalho, foi utilizada uma técnica estatística de detecção chamada de Z-Score Modificado, que é um mecanismo que consegue aferir anomalias em janelas de dados [38]. Estimadores de anomalias que são baseados em média e desvio padrão podem ser afetados por valores extremos ou até por um único valor anormal. Com o objetivo de evitar tal problema, a mediana e o desvio absoluto da mediana (MAD) são empregados no método

Z-Score modificado (em inglês, *Modified Z-Score* - MZS) [38]. Desta forma, a métrica é definida por,

$$M_i = \left| \frac{0.6475(x - \tilde{x})}{MAD} \right|, \tag{4.1}$$

em que M_i é o valor da Z-Score, $\tilde{x}$ é a mediana, e MAD é o desvio absoluto da mediana. Valores acima desse limiar, são considerados anômalos. Segundo [20], observações são rotuladas como *outliers*(valores que fogem da normalidade) quando $|M_i| > 3{,}5$. Esse limiar foi determinado através de simulações baseadas em dados pseudo-normais para tamanhos de amostra de 10, 20 e 40.

Para o nosso experimento foi definida uma janela de dados de tamanho 30 para o cálculo do Z-Score, seguindo outros trabalhos que têm por base esse tamanho ao utilizar técnicas estatísticas [5].

4.2.2 Similaridade Entre Sensores

Devido às características aleatórias dos surgimentos de falhas muitas anomalias podem surgir na mesma janela mas em quantidades as vezes distintas ou em leituras distintas, limitando assim a forma de apontar similaridade entre sensores, seguindo o padrão da presente abordagem. Dado isso, para este componente, dois tipos de medidas de similaridade foram utilizadas. Uma delas foi adaptada pelo trabalho às características presentes ao surgimento de dados falhos, e a outra foi utilizada com a sua definição usual. Assim temos: Distância de Chebyshev Adaptada e Correlação de Pearson.

Distância de Chebyshev Adaptada

A distância de Chebyshev é uma métrica definida em um espaço de vetores onde a distância entre dois vetores é a maior de suas diferenças entre suas dimensões de coordenadas . E ela pode ser definida por:

$$\text{Dchebyshev}\,(x, i) = \max_i(|\,x_{\,i} - y_i|), \tag{4.2}$$

Devido à possibilidade da existência de anomalias em leituras diferentes mas na mesma janela em cada sensor, apenas o valor máximo não pode ser considerado para determinar a

similaridade ou diferença entre duas janelas de sensores. Desta forma, o presente trabalho produziu uma adaptação a medida de similaridade. Assim, a regra estabelecida foi:

$$\text{Dchebyshev\_Adaptado}\,(x, y) = |\max_i(x_i - y_i) - \min_i(x_i - y_i)|, \tag{4.3}$$

em que, quando o valor de subtração entre o valor máximo das diferenças e o valor mínimo das diferenças de cada leitura dos sensores for maior que um limiar pré estabelecido, então os dois vetores são considerados como não similares. Na avaliação realizada para este trabalho, o limiar estabelecido foi de valor 15, devido as características dos dados trabalhados nos experimentos. Este Limiar pode ser ajustado a depender das características dos dados a serem avaliados, levando em consideração qual a diferença considerada fora do padrão nesse comparativo de sensores.

Correlação de Pearson

Afim também de avaliar o processo com alguma técnica já estabelecida e sem adaptações, a correlação de Pearson também foi utilizada. Ela é definida da seguinte forma:

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2(y_i - \overline{y})^2}}. \tag{4.4}$$

O coeficiente de Pearson é obtido por meio de um ajuste de mínimos quadrados e um valor igual a 1 representa uma relação positiva perfeita, enquanto que -1 representa uma relação negativa perfeita e 0 indica a ausência de uma relação entre as variáveis. Assim, se o valor absoluto da correlação for menor que 0.3, então os dois conjuntos de dados são considerados como não similares.

4.3 Detecção e Categorização - Segunda Camada

A segunda etapa se estabelece pelo fato de que apesar do mecanismo descrito na Seção 4.1 direcionar o encontro de eventos e de sensores falhos, apenas ele não é suficiente para cobrir todos os possíveis cenários de anomalias de nó. Isso porque vizinhanças que por algum fator aleatório possuem metade ou a maioria de sensores falhos podem induzir

o sistema a considerar nós corretos como falhos e, assim, uma simples comparação entre sensores não é suficiente.

Para isso, um modelo de previsão de dados é utilizado no nó sorvedouro (em nosso caso o ARIMA), para que através do histórico de medições de cada sensor, seja possível determinar se o conjunto de dados do sensor no período que está sendo analisado está seguindo o comportamento esperado, e aliado a fatores de similaridade já identificados na Etapa 1, definir através de um sistema de lógica fuzzy se um nó é confiável ou não.

O funcionamento detalhado da segunda etapa é estabelecido nos seguintes passos:

1. Inicialmente, o *Cluster Head* envia as informações ao sorvedouro através de pacotes de dados. Tais pacotes podem conter as seguintes informações:

   (a) O dado sensoreado com seu referente *id*;

   (b) Caso o processo da primeira etapa tenha se estabelecido, informações sobre o alerta que foi gerado (evento ou normalidade ou probabilidade de nós falhos). No caso de probabilidade de nós falhos, é enviado o nível de similaridade dos sensores entre si.

2. Após receber os dados com informações do sensoreamento e/ou dos elementos do mecanismo da primeira etapa, esses dados são armazenados em uma base de dados;

3. Após o armazenamento, a informação chegada é checada e analisada a fim de saber que tipo de alerta foi gerado na primeira etapa. Caso seja apenas algum alerta de evento ou de normalidade, nenhuma atitude é tomada, porém se for apontada a provável existência de um nó falho, então o classificador é chamado;

4. O elemento classificador então solicita o histórico de dados do sensor à base de dados, desconsiderando a janela que será analisada;

5. A base de dados então envia as informações e para as lacunas deixadas pelos pacotes perdidos se é substituído o valor nulo pela mediana do conjunto de dados. A partir disto, através de um modelo de predição, o classificador gera uma possível janela de dados com elementos baseados no histórico, permitindo a comparação com a janela de dados que efetivamente foi registrada;

6. É feita então a comparação de cada janela de cada sensor, com a janela gerada pelo preditor para assim analisar se a janela atual segue o padrão de seu histórico, gerando um valor de similaridade que servirá como entrada ao sistema fuzzy.

7. Feito isto, as informações de cada sensor em relação a seu histórico são agregadas e enviadas ao componente do sistema fuzzy;

8. Após isto, as informações referentes à similaridade de cada sensor com seus pares são solicitadas à base de dados, pelo componente do sistema fuzzy;

9. A base de dados então envia as informações sobre similaridade entre os sensores, que juntamente com os outros elementos, alimenta o sistema;

10. Dado o sistema implementado, e todos os parâmetros instanciados a inferência fuzzy é aplicada definindo o nível de confiabilidade do sensor, gerando assim três tipos de resposta ao nó: Nó confiável, Nó não confiável e Não-Definido.

A Figura 4.3 apresenta um diagrama que representa a segunda etapa da abordagem proposta.

4.4 Componentes da Abordagem - Segunda Camada

4.4.1 Classificador/Preditor

Nesta segunda etapa da abordagem, existe o elemento preditor e classificador que irá, baseado no histórico dos dados, apontar quais dados deveriam ser os prováveis dados para as janelas daquele período que está sendo analisado e comparado. Para isso, o método de predição utilizado foi um algoritmo de Séries Temporais, denominado ARIMA (*Autoregressive Integrated Moving Average*).

O ARIMA é um um modelo de previsão de séries temporais baseado em dois conceito principais: autocorrelação e médias móveis [62]. Tal modelo foi desenvolvido em meados dos anos 70, na tentativa de descrever as mudanças que ocorrem em uma série temporal, utilizando uma abordagem matemática. O modelo busca ajuste de valores observados, visando reduzir a diferença dos valores gerados pelo modelo e os valores reais observados.

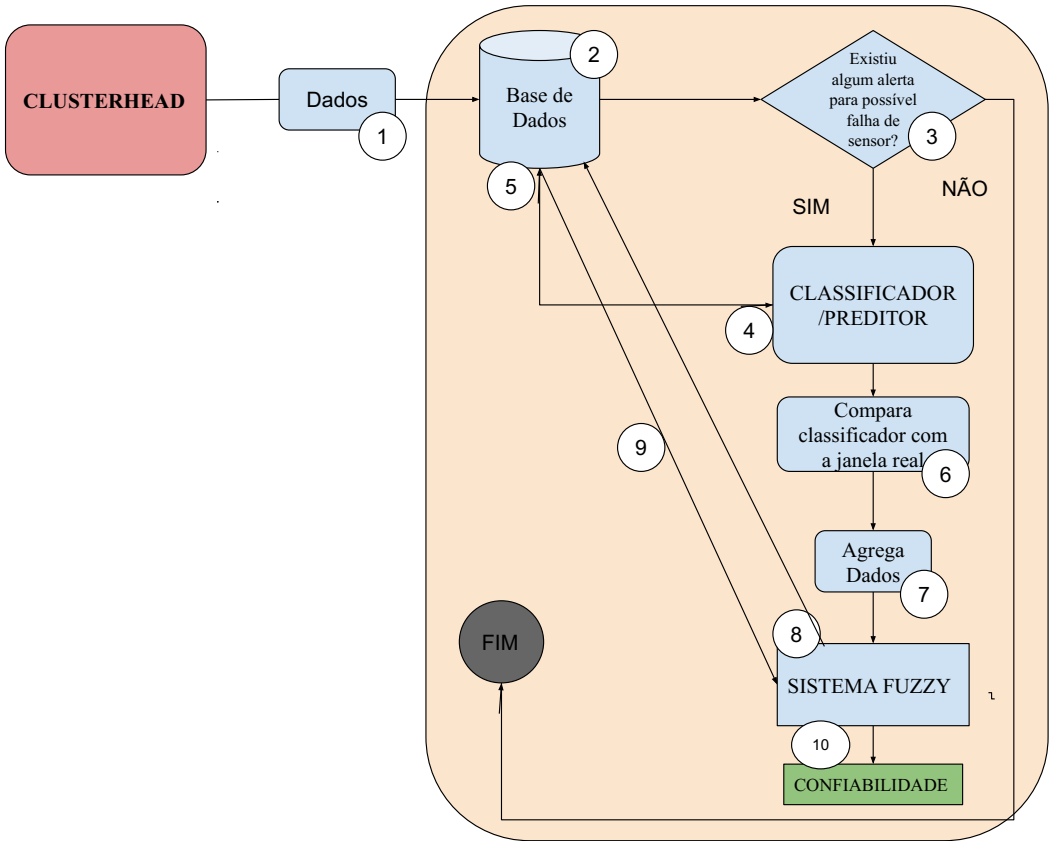

Figura 4.3: Diagrama da abordagem de categorização da segunda etapa.

Além disso, o ARIMA possui a possibilidade de descrever comportamento de séries estacionárias e não estacionarias, permitindo uma maior versatilidade para uma gama variada de situações [29][62].

O modelo ARIMA é composto por três elementos: o modelo auto regressivo (*AutoRegressive Model* (AR)), a Integração (*Integration*(I)) e o modelo de média móvel (*Moving Average Model* (MA)). O modelo AR opera sob a premissa de que os valores passados afetam os valores atuais. O modelo MA assume que o valor da variável dependente no período atual depende dos termos de erro dos períodos anteriores. E o modelo I adiciona diferenciação, subtraindo o valor atual do anterior e podendo ser usada para transformar uma série temporal em uma série estacionária [29][62].

Assim, três parâmetros (p, d, q) são normalmente usados para parametrizar os modelos ARIMA, como descrito a seguir:

- p: número de termos autorregressivos (pedido AR)

- d: número de diferenças não sazonais (ordem diferencial (I))
- q: número de termos de média móvel (ordem MA)

Para a definição desses parâmetros, foi utilizada a função *auto_arima()* da biblioteca *pmdarima* [1], que foi utilizada para a definição dos parâmetros de predição.

4.4.2 Inferência Fuzzy

Como explicitado na Seção 2.4, para o funcionamento do sistema fuzzy são necessárias algumas definições e elementos. Em específico, para o processo de inferência fuzzy dois componentes são importantes: as funções de pertinência e as regras.

Funções de Pertinência

Em nosso sistema fuzzy três componentes considerados antecedentes são definido para estabelecer o nível de confiança que se tem sobre um nó sensor específico: a quantidade de sensores da vizinhança que não são similares a ele, a comparação dos dados da janela recente com seu histórico e a relação dos sensores diferentes comparados ao seus respectivos dados históricos.

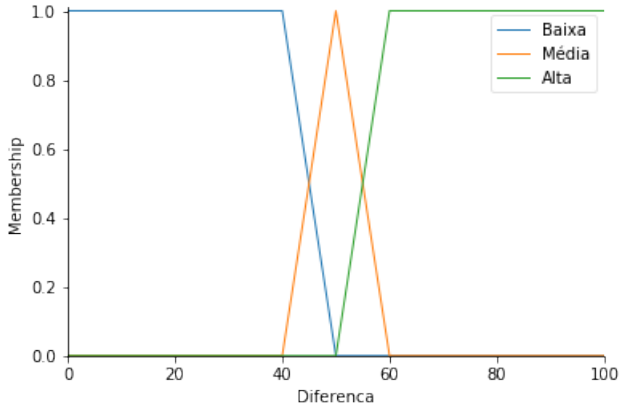

Figura 4.4: Função de pertinência da diferença entre sensores.

As Figuras 4.4, 4.5 e 4.6 apresentam as funções de pertinência dos elementos antecedentes, para a inferência do sistema fuzzy da abordagem. No que consiste a similaridade do

[1] https://pypi.org/project/pmdarima/

sensor com sua vizinhança, um nó sensor pode ter uma semelhança baixa, em que o mesmo é diferente de menos da metade dos vizinhos existentes, média no qual a vizinhança em relação a semelhança com sensor analisado se divide, e alta quando mais da metade dos respectivos vizinhos são diferentes. As funções utilizadas para esse elemento antecedente foram funções triangulares e trapezoidais, e estão dispostas na Figura 4.4.

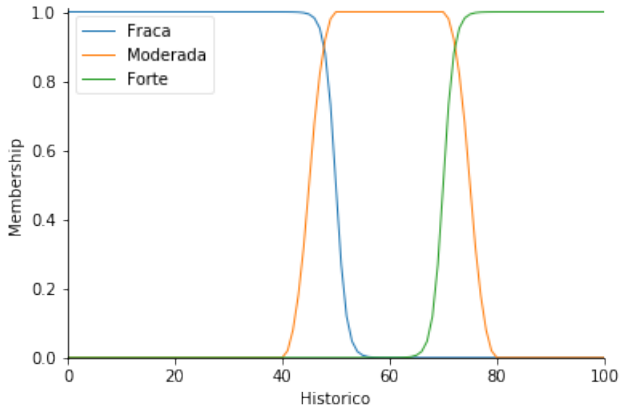

Figura 4.5: Função de pertinência da correlação do sensor com seu histórico.

A janela de dados momentânea do nó sensor pode estar semelhante à janela gerada pelo modelo de predição baseado no histórico dos dados ou não. Com o objetivo de estabelecer o nível de concordância entre a janela de dados e o histórico, a correlação de pearson foi escolhida para determinar está equivalência. Isso ocorre, devido ao fato desse coeficiente já possuir limiares estabelecidos na literatura, que podem assim determinar a afinidade entre os grupos dos dados. Em específico, consegue estabelecer em que nível de semelhança está os dados da janela analisada com os dados gerados pelo modelo de predição baseado no histórico.

A Figura 4.5 ilustra como a função de pertinência foi estabelecida para este caso. No qual o conjunto de dados em análise do sensor, pode se considerado de forma fraca, onde não há similaridade entre os dados da janela com seus dados passados, moderada onde não é possível estabelecer com certeza essa similaridade ou forte quando os dados estão de acordo com seu histórico, gerando a informação que se é esperada para aquele momento.

A ultima etapa está relacionada aos vizinhos que foram considerados diferentes do sensor em análise. Os vizinhos considerados diferentes do nó sensor podem ser em sua maioria correlacionados com os seus respectivos históricos, podem ser em sua maioria não correla-

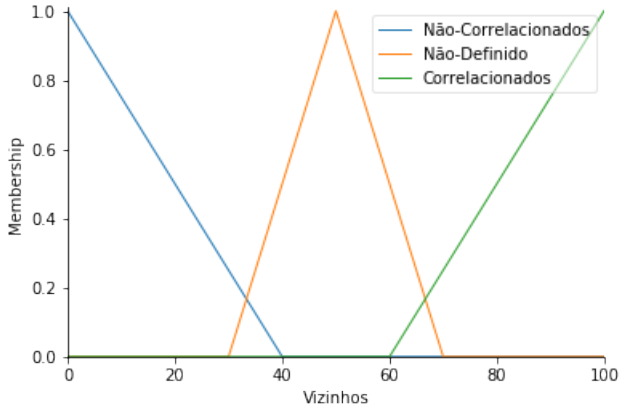

Figura 4.6: Função de pertinência da correlação dos vizinhos diferentes com seus históricos.

cionados, e por fim pode existir parte dos vizinhos diferentes com convergência com seus históricos e parte com características divergentes dos seus dados anteriores. Foram utilizados funções triangulares para estabelecer como são enquadrados os termos referentes a correlação dos vizinhos, como explicitado na Figura 4.6.

Já o elemento consequente (resposta aos antecedentes), tem por objetivo atribuir um grau de confiança ao nó sensor avaliado, em consequência dos valores anteriores. E com isso, os estados possíveis para um determinado sensor pode ser definido da seguinte maneira:

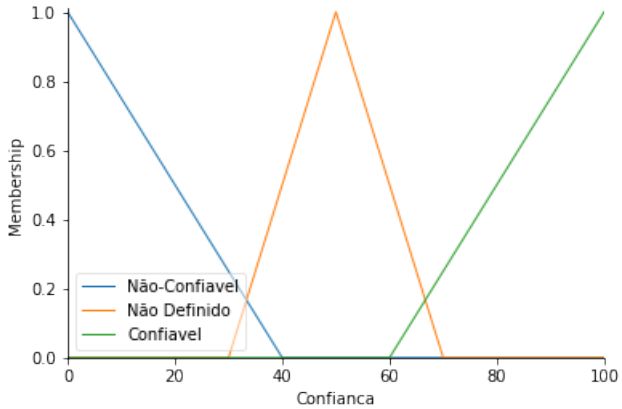

Figura 4.7: Função de pertinência consequente que define a confiança dos sensores.

- Não-Definido: Quando o modelo não conseguiu definir se o nó sensor é falho ou não; Isso ocorre quando os valores estabelecidos na relação do nó com seu histórico e com seus vizinhos não foram suficientes para que o modelo determine se há confiabilidade ou não no sensor.

- Não-Confiável: Quando o modelo definiu que o nó sensor não é confiável e provavelmente é falho. Isso ocorre quando o sensor se apresenta em desacordo com seu histórico e/ou diferente dos seus vizinhos.

- Confiável: Quando o modelo define que o nó sensor é seguro, e provavelmente está monitorando e enviando os dados corretamente. Isso ocorre quando o sensor está em acordo com seu histórico e/ou com seus vizinhos.

As funções de pertinência que determinam o estado final do sensor, são explicitadas na Figura 4.7. E foram construídas baseadas em funções triangulares e trapezoidais.

Regras

As regras que determinam o funcionamento do sistema fuzzy são definidas de acordo com a Tabela 4.1.

Regra	**Diferença**	**Histórico**	**Vizinhos**	**Confiança**
1	Alta	Fraco	Correlacionados	Não-Confiável
2	Baixa	Forte	Correlacionados	Confiável
3	Média	Moderado	Não-Definido	Não Definido
4	Alta	Forte	Não-Correlacionados	Confiável
5	Baixa	Fraca	Não-Correlacionados	Não-Confiável
6	Alta	Fraca	Não-Correlacionados	Não Definido
7	Alta	Fraca	Não-Definido	Não-Confiável
8	Média	Fraca	Não-Correlacionados	Não Definido
9	Baixa	Forte	Não-Correlacionados	Confiável
10	Média	Forte	Não-Correlacionados	Confiável
11	Média	Fraca	Correlacionados	Não-Confiável
12	Alta	Forte	Não-Definido	Confiável
13	Baixa	Fraca	Não-Definido	Não Definido

Tabela 4.1: Regras de inferência fuzzy.

O sistema fuzzy foi estabelecido a partir da biblioteca scikit-fuzzy [2] do python.

[2] http://pythonhosted.org/scikit-fuzzy/

Capítulo 5

Avaliação da Proposta

Neste capítulo é exposto o processo para avaliação da solução proposta. O objetivo da avaliação é averiguar o funcionamento do mecanismo desenvolvido na detecção e categorização de anomalias em Redes de Sensores sem Fio (RSSFs). O planejamento foi conduzido a fim de responder às seguintes questões de pesquisa:

- **Questão de pesquisa 1 - QP1**: A abordagem consegue categorizar as anomalias de maneira eficiente?
- **Questão de Pesquisa 2 - QP2**: No contexto de falhas nos dados, a abordagem consegue categorizar falhas de diferentes formatos nos sensores com problemas?
- **Questão de Pesquisa 3 - QP3**: A abordagem consegue determinar a existência de sensores falhos à medida que em que eles vão se tornando maioria na vizinhança?
- **Questão de Pesquisa 4 - QP4:** A abordagem sofre alguma redução na eficiência de detecção devido às perdas de dados ocorridas na rede?

5.1 Ambiente de simulação

Para a construção das simulações dos cenários de RSSF, o *framework* Castalia 3.0 foi o escolhido. O Castalia é um simulador de código aberto desenvolvido com a plataforma OMNeT++ para RSSF e *Body Area Networks* (BANs) [8].

A escolha de tal ferramenta se motiva pela grande aceitação e uso na comunidade de RSSF, devido às características realísticas para simulação de comportamentos de meios de transmissão, modelos de rádio, modelos de bateria e até simulações dos processos físicos dos sensores. Além disso, o simulador possibilita que os desenvolvedores criem seus próprios algoritmos e protocolos.

As simulações neste trabalho foram realizadas em um computador com sistema operacional ubuntu 16.04, processador Intel Core i5-4200U como 8 núcleos, 16GB de memória RAM e 1TB de disco.

5.2 Estudo de Caso

Dadas as inúmeras aplicações existentes no contexto das RSSFs, foi selecionado o ambiente industrial como um contexto relevante para a averiguação do mecanismo proposto. Para isso são necessárias algumas configurações no canal sem fio que estabeleça o padrão de comunicação existente nesse tipo de aplicação. E assim, baseado nos conceitos estabelecidos nos trabalhos descritos em [46][17], o script de simulação do Castalia foi configurado seguindo parâmetros mostrados na Tabela 5.1.

SN.wirelessChannel.pathLossExponent = 1.69
SN.wirelessChannel.PLd0 = 80.48
SN.wirelessChannel.d0 = 15
SN.wirelessChannel.sigma = 6.62
SN.wirelessChannel.K = 12.3
SN.wirelessChannel.K_sigma = 5.4
SN.wirelessChannel.meanTimeChange = 85
SN.wirelessChannel.seed = 0

Tabela 5.1: Parâmetros de configuração do canal sem fio para redes industriais.

Os quatro primeiros parâmetros calculam a perda de percurso e sombreamento, enquanto que os parâmetros K e K_σ são usados para calcular a atenuação em pequena escala. O parâmetro *meanTimeChange* aponta para o tempo médio em que uma mudança ocorre na características dos canais, em minutos e o parâmetro *seed* define a semente que gera valores

aleatórios de potência recebida durante a simulação [46] [17].

Para parâmetros ligados aos modelos de sombreamento log-normal e Rice, foram utilizados dados obtidos de experimentos de ambiente industrial, que consideram um cenário sem visada direta entre transmissor e receptor. O nível de potência de transmissão foi de 0 dbm. A taxa de transmissão de pacotes foi de 0,2 pacotes/s, seguindo a lógica de transmissão de 1 pacote a cada cinco segundos. Também foi empregado o protocolo CSMA/CA na camada MAC, definido no padrão IEEE 802.15.4. A Tabela 5.2 apresenta em maior detalhes os parâmetros da simulação para a configuração completa de um ambiente industrial [46] [17].

Camada física e MAC	IEEE 802.15.4 - CSMA/CA
Taxa de bits	250 kbit/s
Potência de transmissão	0 dBm
Taxa de transmissão de pacotes	0,2 pacotes/s
Tempo médio de mudança (Tc)	85 minutos
Expoente de perda de percurso (n)	1,69
Distância de referência (d0)	15 metros
Perda de percurso na distância de referência (L(d0))	80,48 dB
Desvio padrão do sombreamento (X)	8,13 dB
Fator de Rice (K)	12,3 dB
Desvio padrão do fator de Rice (K)	5,4 dB

Tabela 5.2: Parâmetros do script do castalia de configuração do ambiente de uma rede industrial.

5.3 Base de Dados

Para o estudo, foram consideradas leituras históricas de sensores coletadas de 54 sensores implantados no laboratório da *Intel Berkeley Research* entre 28 de fevereiro e 5 de abril de 2004. O conjunto de dados contém 2,3 milhões de leituras de sensores. Esses sensores coletaram valores de hora, temperatura, umidade, luminosidade, tensão e época, que é um índice demarca um período no qual os sensores coletaram a informação. Tudo isso coletado por meio do sistema de processamento de consultas *TinyDB* na rede [36], construído na

plataforma *TinyOS* [1]. A Figura 5.1 apresenta a distribuição dos sensores do Laboratório *Intel Berkeley* [36].

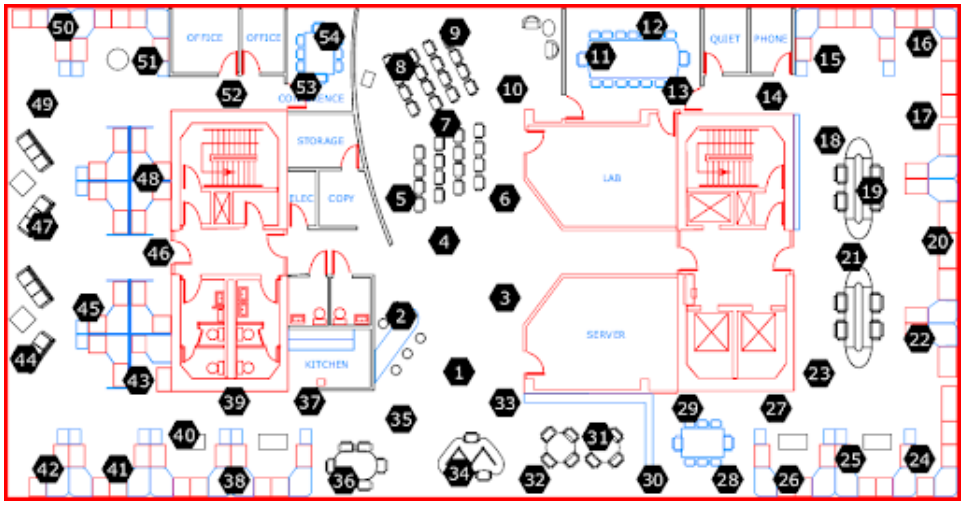

Figura 5.1: Ilustração do Laborátorio Intel Berkeley com os Sensores

.

5.3.1 Organização e Pré-Processamento dos Dados

Previamente à implementação do ambiente de experimentação, os dados do conjunto a ser analisado foram pré-processados. Afim de reduzir tempo de simulação e processamento, para os experimentos foram considerados os dados referentes a temperatura no período entre 28 de fevereiro e 13 de março. A partir deste conjunto de dados, uma limpeza foi realizada retirando informações duplicadas, incompletas e que não condiziam com a estrutura de dados do arquivo.

Após este pré-processamento, foi necessário estabelecer as vizinhanças que iriam compor o sistema hierárquico da abordagem e como os sensores iriam ser distribuídos. Para isso, seguindo o trabalho descrito em [51], uma técnica de clusterização *KNN* foi aplicada, definindo as vizinhanças de acordo com a Figura 5.2.

Apesar do trabalho descrito em [51] ter definido os *Cluster Heads*, como a Figura 5.2 demonstra, o critério para definir o nó agregador(CH) neste trabalho foi diferente. Em geral procurou-se definir como *Cluster Head* os sensores que possuíssem menos dados ou os dados

[1] https://github.com/tinyos/tinyos-main

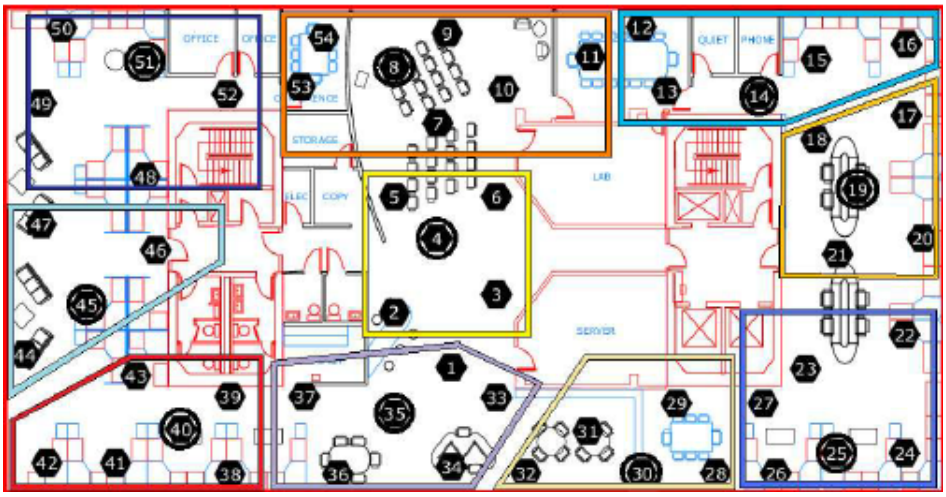

Figura 5.2: lustração do Laborátorio Intel Berkeley com as Vizinhanças estabelecidas

.

mais danificados e incoerentes, já que em nosso experimento o *Cluster Head* não tem o sensoriamento como função. Além disso, devido à estrutura hierárquica e a ausência de um nó sorvedouro, que deve ser criado, as posições dos nós foram ajustadas de maneira aleatória, utilizando a biblioteca Networkx[2] do python, respeitando a divisão dos nós em suas respectivas vizinhanças, de forma a organizar os dados de cada *cluster* de acordo com a região monitorada do laboratório.

Também, devido à variação na quantidade de dados de cada sensor, mesmo compondo a mesma vizinhança, com o objetivo de nivelar essa quantidade de dados, e manter os dados dos sensores mais similares possíveis, só foram considerados dados dos sensores que coletaram as informações na mesma época, seguindo o valor do índice *Epoch*. Este índice como já explicitado, ele demarca momentos em que os sensores enviam as informações coletadas, possibilitando desta forma, estabelecendo este critério os dados das vizinhanças se tornam similares e equiparáveis.

É importante destacar que mesmo após o pré-processamento dos dados, a vizinhança composta pelos sensores 17,18,19,20,21 não apresentou a similaridade prevista, apresentando dados discrepantes entre si. Devido a dificuldade de alocação desses sensores em outras vizinhança, este *cluster* em específico foi descartado, mantendo-se apenas 9 vizinhanças ao total. A Tabela 5.3 demonstra como ficaram distribuídos os sensores nas vizinhanças.

[2]https://networkx.org/

Vizinhança	Cluster Head	Sensores
1	5	2,3,4,6
2	54	7,9,10,11,53,8
3	15	12,13,14,16
4	24	22,23,25,26,27
5	28	29,30,31,32
6	35	1,33,34,36,37
7	40	38,39,41,42,43
8	45	44,46,47
9	51	48,49,50,52

Tabela 5.3: Distribuição de sensores por vizinhança.

5.3.2 Criação de Anomalias Sintéticas

A base de dados utilizada não possui anomalias rotuladas. Devido a isso, após o pré-processamento descrito na Seção 5.3.1, um mecanismo de geração de anomalias sintéticas foi construído a fim de representar anomalias do mundo real, e assim permitir que a abordagem consiga ser analisada.

Para tal, utilizou-se uma abordagem que se baseia em uma cadeia de Markov de dois estados. Uma cadeia de Markov é um processo estocástico de Markov que toma valores inteiros [17]. Tal processo só pode ser assim denominado se satisfizer a propriedade de Markov, que afirma que a probabilidade do estado do processo em um instante k+1 depende somente do estado processo no instante k.

A Figura 5.3 ilustra os dois estados da cadeia de Markov utilizada neste trabalho para geração das anomalias sintéticas.

Enquanto a cadeia de Markov permanece no estado Normal, os dados permanecem inalterados, ou seja com o mesmo valor do *dataset*. A transição para o estado de Anomalia ocorre com a probabilidade p, assim, quanto maior essa probabilidade, mais frequente surgem anomalias. A duração do estado de Anomalia está ligado à probabilidade q, quanto menor essa probabilidade menor a chance de alteração do estado de Anomalia para o estado Normal, gerando então rajadas de anomalias maiores, ou seja uma maior concentração de

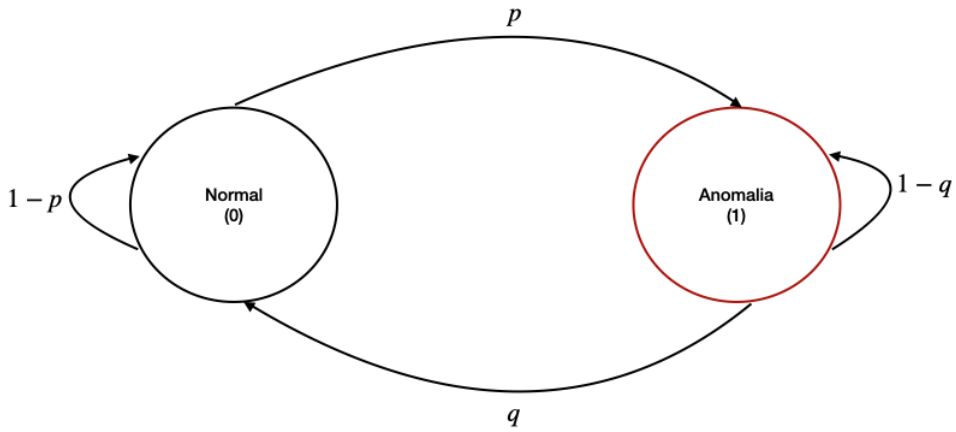

Figura 5.3: Exemplo do funcionamento da Cadeia de Markov na criação de anomalias

.

dados anômalos. Quando o estado de anomalia está estabelecido, o valor real do *dataset* é somado a um valor aleatório gerado por uma distribuição gaussiana, em uma faixa de valores estabelecida, tornando assim o dado fora do padrão e consequentemente anômalo.

5.4 Cenários Avaliados

Para a avaliação da abordagem, quatro cenários foram desenvolvidos, a partir da geração de anomalias sintéticas descritas na Seção 5.3.2.

- **Cenário 1**: O primeiro cenário visa representar o surgimento de anomalias **pontuais** oriundas de falhas de sensor, de maneira **intermitente**, de forma que a cada janela o sensor falho apresente uma ou poucas anormalidades;

- **Cenário 2**: O segundo cenário visa representar o surgimento de anomalias **mais frequentes** oriundas de falhas de sensor, de maneira **intermitente**, de forma que a cada janela o sensor falho apresente anormalidades em maior quantidade;

- **Cenário 3**: O terceiro cenário visa representar o surgimento de falhas de sensor **graduais**, iniciando de forma pontual e evoluindo ao longo do tempo a se tornarem coletivas, de forma que os sensores falhos passem a produzir apenas dados anormais em determinado ponto;

- **Cenário 4**: O quarto cenário tem por objetivo gerar eventos na **vizinhança**, gerando anomalias **equivalentes** em todos os sensores do *cluster* ao mesmo tempo.

A Tabela 5.4 apresenta como ficou disposta a quantidade de anomalias e os valores empregados dos parâmetros p e q para cada cenário.

Cenário	p	q	**Percentual de Anomalias**
Cenário 1	0.05	0.99	1-3%
Cenário 2	0.08	0.6	8-10%
Cenário 3	0.5; 0,8	0.99; 0.6; 0.3; 0.08	40-50%
Cenário 4	0.0145	0.99; 0.7	2-4%

Tabela 5.4: Percentual de anomalias por cenário

A definição da quantidade de nós falhos em uma vizinhança assim como, quais nós serão os falhos é definido de maneira aleatória através da função random da biblioteca numpy do python, que gera valores aleatórios baseados em uma função de probabilidade gaussiana.

5.4.1 Métricas de Avaliação

Com o objetivo de avaliar o funcionamento da abordagem na identificação do estado dos sensores em cada janela analisada de cada vizinhança, três métricas foram avaliadas. A primeira métrica é a acurácia, que é a proximidade de um resultado com o seu valor de referência real. Dessa forma, quanto maior a acurácia, mais próximo da referência ou valor real é o resultado encontrado. A acurácia pode ser definida como,

$$\text{Acurácia} = \frac{(VP + VN)}{(VP + VN + FP + FN)}, \tag{5.1}$$

em que VP é o percentual e verdadeiros positivos, que representa as ocorrências de sensores considerados falhos ou eventos que foram corretamente identificados. Já o VN são os verdadeiros negativos, que representa as não ocorrências de sensores considerados falhos ou eventos corretamente identificados. O FP são os falsos positivos, que representam alarmes errôneos identificados pela abordagem, identificando eventos ou falhas sem os mesmos ocorrerem, em nosso caso também foram considerados falsos positivos alarmes da abordagem estabelecidos como Não-Definidos, vide Seção 4.4.3. Por fim, o FN que são os falsos

negativos que representa a ocorrência de eventos ou sensores falhos não identificados nas janelas de análise.

Outra métrica é o *Recall* ou Taxa de acertos, que é a proporção de verdadeiros positivos, ou seja, a capacidade do sistema em predizer corretamente a condição. Ela é calculada pelo o número de predições positivas corretas dividido pelo número total de positivos. E pode ser definido por,

$$\text{Recall} = \frac{(VP)}{(VP + FN).} \tag{5.2}$$

Por fim, foi calculada a taxa de falsos alarmes, que indica proporção de predições erroneamente calculadas pelo sistema. Ela é calculada pelo o número de falsos positivos dividido pelo número total de verdadeiros negativos, e é definida por

$$TaxadeFalsosAlarmes = \frac{(FP)}{(FP + VN).} \tag{5.3}$$

5.5 Tratamentos

Para o design e configuração do experimento, se foi escolhido o Design Fatorial Completo. Por meio deste design é possível comparar o desempenho dos cenários, considerando as diferentes combinações entre os níveis dos fatores, que são: os cenários estabelecidos e as medidas de similaridade definidas para execução da abordagem. Rodadas preliminares do experimento foram realizadas e, fazendo uso dos resultados, foi calculado o tamanho da amostra, para se obter 90% de confiança. Como resultado definiu-se como 10 o número de repetições de cada um dos tratamentos expressos na Tabela 5.5. Portanto, essa configuração resulta em um total de 80 execuções da simulação. Esse valores de tamanho de amostra e replicações foram estabelecidos considerando o tempo de simulação e processamento dos experimentos.

Tratamento	Medida de Similaridade	Cenário	Repetições
1	Chebyshev Adaptado	1	10
2	Chebyshev Adaptado	2	10
3	Chebyshev Adaptado	3	10
4	Chebyshev Adaptado	4	10
5	Correlação	1	10
6	Correlação	2	10
7	Correlação	3	10
8	Correlação	4	10

Tabela 5.5: Tratamentos

5.6 Ameaças a Validade

Das ameaças à validade constatadas, destaca-se uma ameaça à validade interna ao utilizar o Castalia para simular ambientes RSSFs. É de conhecimento notório que qualquer simulação tem um erro inerente associado, logo, os dados podem sofrer alterações dada a utilização deste ambiente. Entretanto, conforme mencionado na Seção 5.1, o Castalia é uma ferramenta largamente usada em diversas pesquisas, evidenciando possuir um bom nível de qualidade.

Conjuntamente, também é possível observar uma ameaça à validade externa, uma vez que, fazendo uso de ambientes simulados não é confiável generalizar os resultados para cenários reais. Com o intuito de mitigar este fator, foram utilizados dados de bases públicas, coletados de um cenário real. Além disso, como ameaça a validade externa temos o fator da geração de anomalias sintéticas, que só foi realizada devido ao fato de não se existir no *dataset* anomalias rotuladas, para mitigar tal ameaça, as anomalias foram geradas de maneira aleatória utilizando uma distribuição gaussiana, que é a distribuição dos dados do *dataset*. Por fim, para atenuar ameaças à validade de conclusão, buscou-se realizar experimentos prévios, com os quais fosse possível calcular o tamanho da amostra necessá- ria, assim como foi utilizado o bootstrap como método de reamostragem e estimativas com intervalos de confiança de 90%.

Capítulo 6

Resultados

Este capítulo tem como objetivo apresentar os resultados obtidos por meio da execução dos experimentos descritos no Capítulo 5, a fim de expor os efeitos gerados pela solução proposta, considerando os contextos específicos e a busca por respostas às questões de pesquisa levantadas.

6.1 Avaliação das Métricas

As métricas apresentadas e definidas na Seção 5.4.1 possuem por objetivo apresentar, em diferentes óticas, a efetividade da proposta nos contextos já pré-estabelecidos. A fim de definir esses conceitos, os valores médios de cada métrica foram estimados com 90% através do mecanismo de reamostragem *bootstrap*, por meio dos resultados gerados pelos experimentos e suas respectivas replicações.

6.1.1 Acurácia

A acurácia nos permite avaliar a proximidade dos resultados gerados pela abordagem com o valor real. A Figura 6.1 apresenta os valores médios estimados para a acurácia, considerando os quatro cenários e as duas medidas de similaridade utilizadas no componente de comparação da abordagem descrita no Capítulo 4.

Comparando inicialmente a utilização das duas medidas, apenas é possível afirmar que existe diferença no nível de acurácia entre Chebyshev Adaptada (CA) e a Correlação (CO) no

Cenário 1, no qual o valor médio estimado da abordagem ao utilizar o CA está contido entre 82.80% e 87.20%, enquanto que ao utilizar CO estima-se que a acurácia fica entre 78.22% e 82.30%. Dessa forma, pode-se afirmar, com 90% de confiança, que existe uma diferença para este cenário, que pode ser mínima de apenas 0.50% ou de até 9%.

Nos demais cenários, os valores estimados para as duas medidas se interseccionam, não se podendo afirmar a existência de diferença ou não nesses casos.

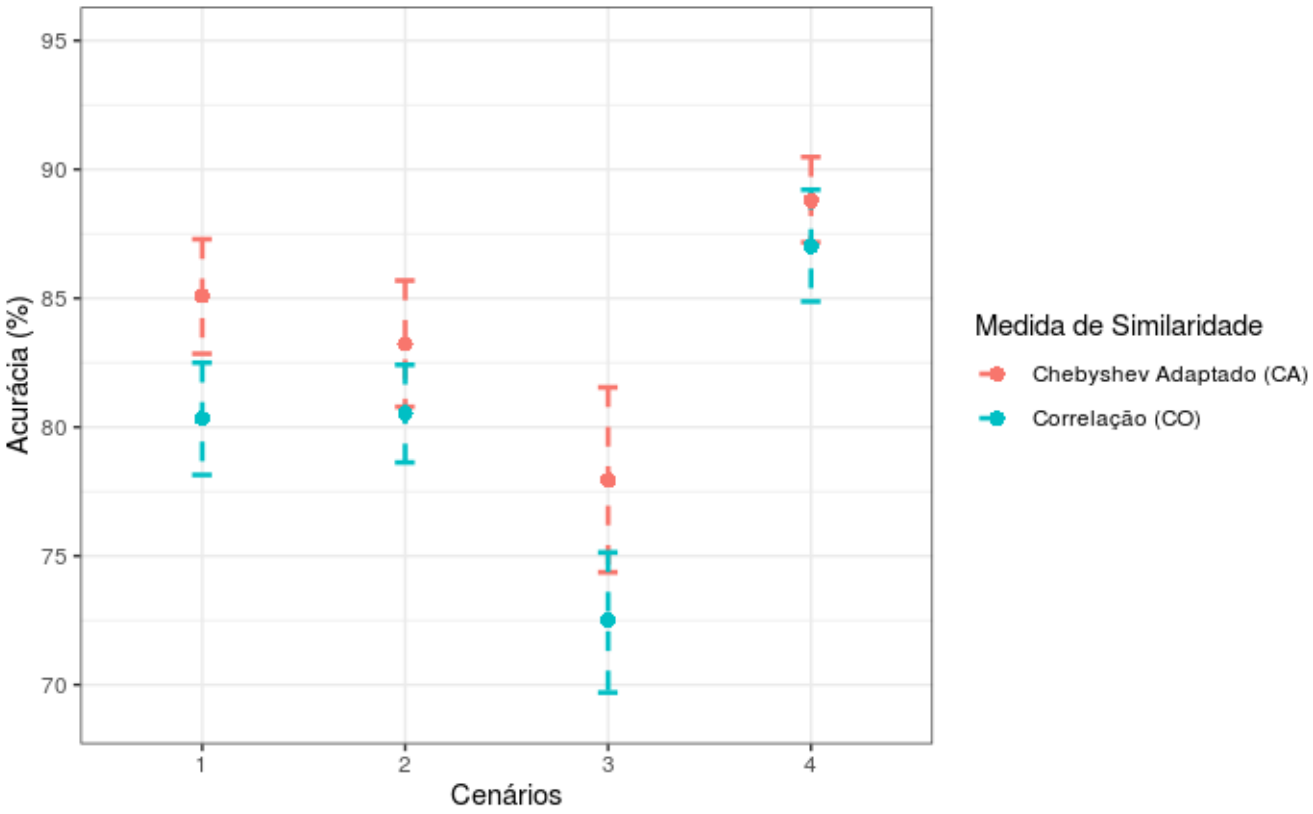

Figura 6.1: Valores médios estimados de Acurácia.

.

No comparativo entre cenários, destaca-se o Cenário 4 com valores de acurácia estimados entre 85.99% e 90.05% para o Chebyshev Adaptado e entre 85.01% e 90.12% para correlação. Apenas em comparativo ao Cenário 1 e no contexto do uso do Chebyshev Adaptado, não se pode afirmar a existência ou não de diferença entre os dois cenários. Considerando os outros cenários, bem como a medida de correlação, nota-se que a acurácia para o cenário de eventos pode ser considerada mais alta, e que isto pode ser ocasionado tanto pela efetividade em detectar os eventos, como pela menor geração de falsos positivos, ou até pelo relação dos

resultados dessas duas métricas, e que serão discutidas nas seções posteriores.

Outro ponto de destaque se dá ao Cenário 3, em especial à medida de Correlação (CO), em que se pode verificar uma menor acurácia em relação aos outros cenários, em que ao utilizar a medida CO o intervalo está estabelecido entre 69% e 75%. No contexto do uso do Chebyshev Adaptado, a diferença é existente em relação aos cenários 1 e 4, e pode existir ou não em relação ao Cenário 2, com valor estimado entre 74% e 81%, enquanto que o Cenário 2 este valor médio presumido está contido entre 80% e 85%. Ressalta-se também a intersecção entre os cenários 1 e 2, não podendo também se afirmar diferença entre estes cenários.

Para mais detalhes, a Tabela 6.1 apresenta a faixa de valores de acurácia média inferidos para cada cenário.

Cenário	**Média - Acurácia**	**Limite Inferior**	**Limite Superior**	**Medida de Similaridade**
1	85.03	82.80	87.20	Chebyshev Adaptado (CA)
2	83.19	80.68	85.67	Chebyshev Adaptado (CA)
3	77.90	74.46	81.18	Chebyshev Adaptado (CA)
4	88.06	87.10	90.05	Chebyshev Adaptado (CA)
1	80.33	78.22	82.30	Correlação (CO)
2	80.52	78.59	82.37	Correlação (CO)
3	72.51	69.66	75.38	Correlação (CO)
4	87.66	84.86	89.17	Correlação (CO)

Tabela 6.1: Valores Estimados de Acurácia

As diferenças entre os valores de acurácia, considerando cenários e medidas, em geral estão relacionadas ou a capacidade de identificação correta de sensores falhos e eventos, ou da geração de falsos alarmes, para entender melhor estes valores e o que gera, com maior direcionamento, essa taxa de efetividade geral, a análise de recall e da taxa de falsos alarmes

é necessária, e serão discutidas nas subseções seguintes.

6.1.2 Recall

O recall tem por objetivo em nosso contexto, conseguir apontar a efetividade da abordagem em identificar corretamente nós sensores falhos e eventos, em seus respectivos cenários. Com essa informação, podemos então analisar em que nível de acerto em relação as anomalias, o mecanismo proposto está. Considerando tal métrica, a Figura 6.3, apresenta os valores médios estimados de recall para os cenários e medidas de similaridade utilizados nos experimentos.

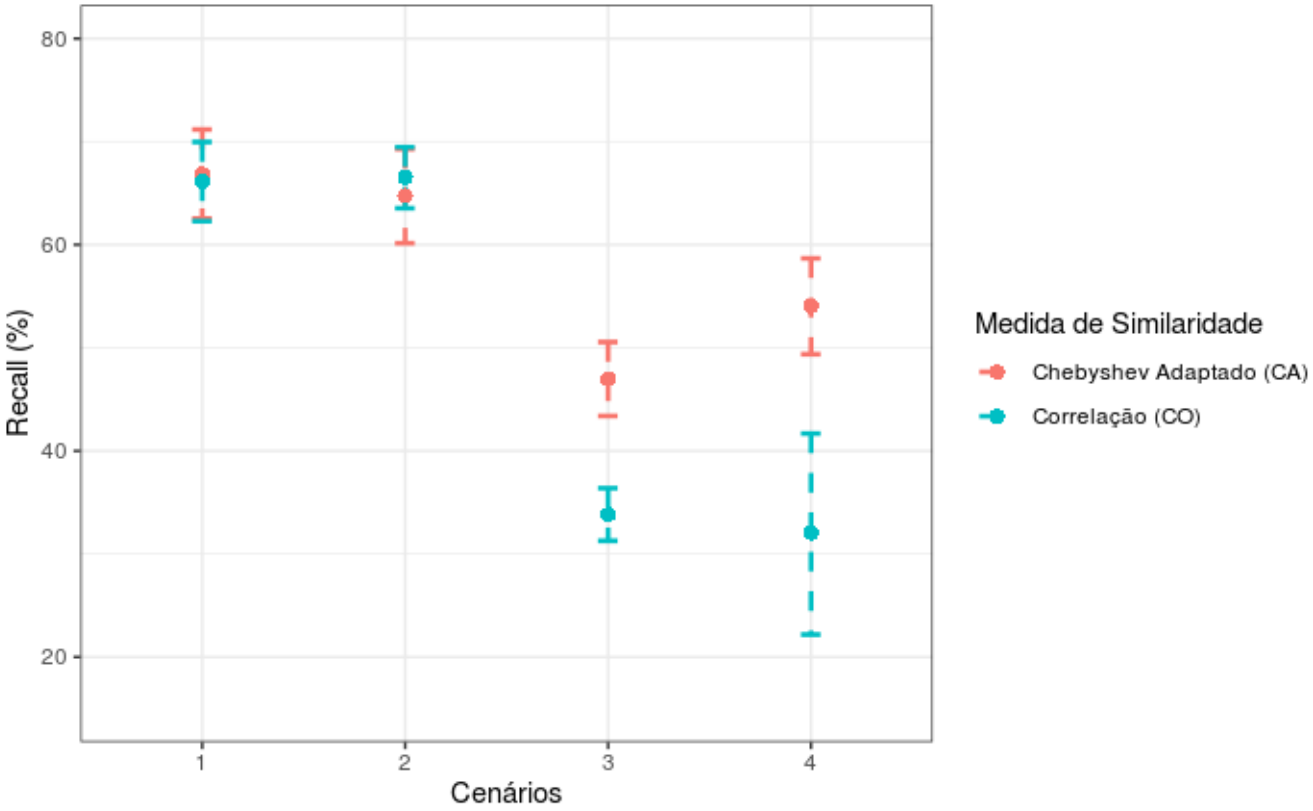

Figura 6.2: Valores médios estimados de Recall.

Considerando inicialmente as medidas de similaridade, percebe-se que em relação aos cenários, os grupos se dividem em dois. Nos cenários 1 e 2, não é possível afirmar a existência de diferença ou não entre as medidas utilizadas, com intervalos que se interseccionam. Já considerando os cenários 3 e 4, essa lógica se torna diferente, é possível afirmar que nos

dois cenários existe diferença entre a taxa de encontro de sensores falhos ou eventos, com valores estimados maiores para a medida Chebyshev Adaptado.

No Cenário 3, o valor de Recall para o CA está estimado entre 43% e 50%, enquanto que para a CO, estima-se o valor médio de Recall entre 31% e 36%, sendo possível afirmar a existência de uma maior efetividade para o Chebyshev Adaptado, com uma diferença que pode variar de 7% a 19%.

Já considerando o Cenário 4, nota-se também uma diferença entre as duas medidas, em que se estima que o valor de Recall está concentrado entre 49% e 58% para Chebyshev Adaptado, enquanto que para Correlação o valor estimado se concentra entre 22% e 41%, podendo assim haver uma diferença de 8% a 36%.

Cenario	**Média - Recall**	**Limite Inferior**	**Limite Superior**	**Medida de Similaridade**
1	66.84	62.32	71.21	Chebyshev Adaptado (CA)
2	64.75	59.94	69.42	Chebyshev Adaptado (CA)
3	46.98	43.35	50.69	Chebyshev Adaptado (CA)
4	54.06	49.23	58.69	Chebyshev Adaptado (CA)
1	66.16	62.27	69.98	Correlação (CO)
2	66.53	63.50	69.26	Correlação (CO)
3	33.81	31.04	36.27	Correlação (CO)
4	32.05	22.37	41.76	Correlação (CO)

Tabela 6.2: Valores Estimados de Recall

Considerando o comparativo entre cenários, é perceptível que nos cenários 1 e 2 houve uma melhor resposta na identificação das anomalias e do estado dos sensores, em comparação com os cenários 3 e 4. As falhas intermitentes ocorridas nos sensores, sejam mais ou menos frequentes, induziram mais facilmente ao sistema identificar de forma efetiva a ocorrência dos erros nos sensores. Em compensação, nos cenários 3 e 4 a efetividade foi

menor.

A avaliação realizada no Cenário 3 demonstra que, com a evolução da quantidade de anomalias nos sensores falhos, os dados considerados anormais tendem a começar a serem considerados normais pelo elemento de detecção, o M-ZScore. A técnica acaba não conseguindo identificar as anormalidades, ou apenas identifica de maneira mais esporádica e desajustada, visto que os dados considerados anteriormente anômalos se tornam o padrão dos dados. Somado ainda, ao fato de que o elemento de predição (neste caso o ARIMA) também com o tempo acaba considerando os *outliers* como o histórico padrão dos dados. Isso acaba minando a eficiência na detecção dessas anomalias. Além disso, medidas como a de Correlação, que analisam as tendências dos valores dos conjuntos de dados, acabam tratando sensores diferentes como similares.

Esse aspecto também afeta o cenário de eventos, principalmente para medida de Correlação, em que neste contexto, as oscilações naturais dos dados fazem com que as comparações sejam distorcidas, e por não haver mais um elemento como sistema *fuzzy* para atenuar este problema, os resultados de Recall acabam sendo prejudicados.

Para mais detalhes, a Tabela 6.2 apresenta a faixa de valores de acurácia média inferidos para cada cenário.

6.1.3 Taxa de Falsos Alarmes

Considerando inicialmente as medidas de similaridade, os 1 e 2 e apresentam taxas de falsos alarmes diferentes. No Cenário 1 presume-se que o valor médio da taxa de falsos alarmes está contido entre 8 e 11%, para o Chebyshev Adaptado, enquanto que para Correlação o resultado varia entre 12% e 15%. Enquanto que no Cenário 2 estima-se que o valor médio na utilização do CA está contido entre 7% e 10%, e entre 13 e 16% quando se é utilizado a medida CO.

Já nos cenários 3 e 4, não se pode afirmar diferença ou não entre a utilização das medidas de similaridades no que diz respeito à geração de falsos alarmes, uma vez que nestes três cenários os intervalos de confiança se interseccionam.

Considerando os cenários, alguns pontos merecem destaque. Em relação à utilização da medida de Correlação, os três cenários voltados à existência de nós falhos apresentam intervalos similares, não sendo possível, assim, afirmar a existência de diferença ou não

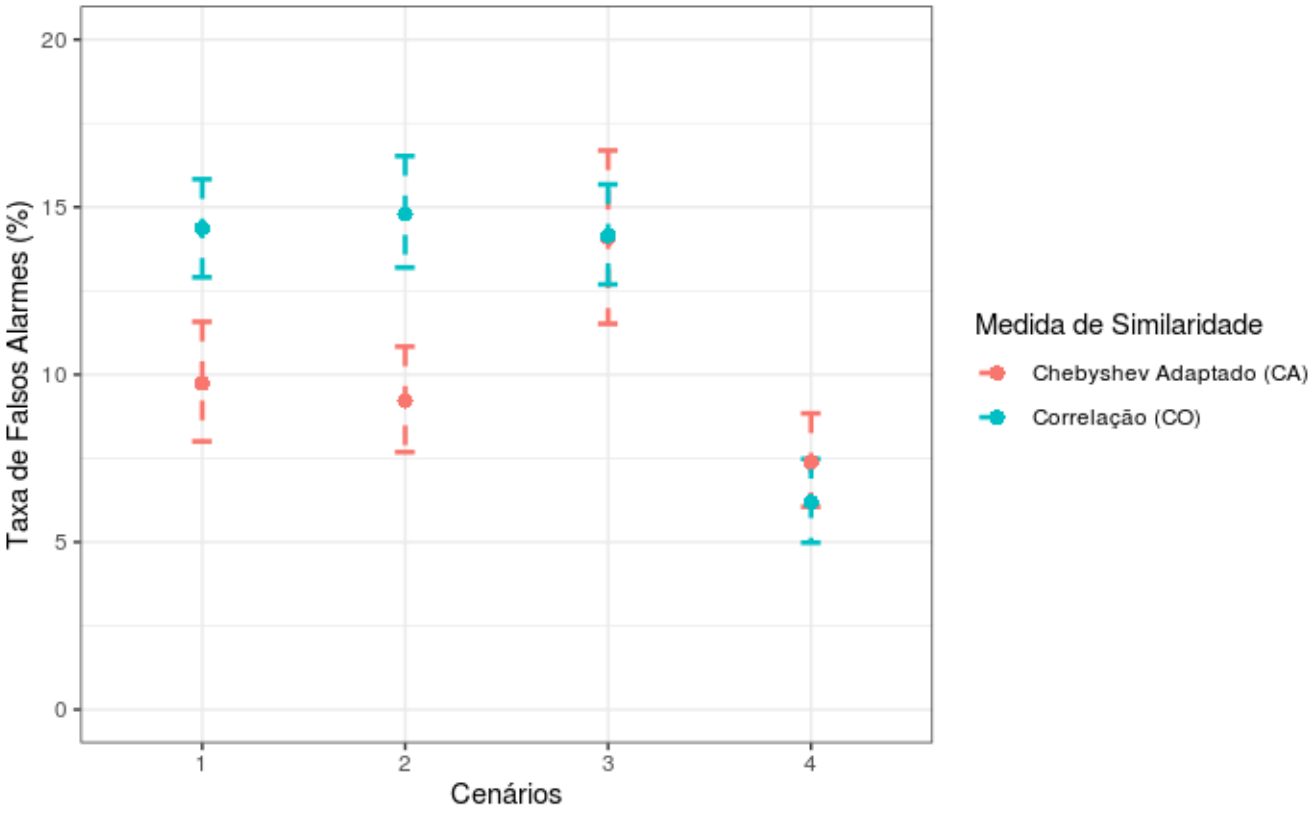

Figura 6.3: Valores médios estimados da Taxa de Falsos Alarmes.

entre eles. Para o Chebyshev, isso ocorre apenas entre os cenários 1 e 2, enquanto que o cenário 3 apresenta diferença em relação ao cenário 2, visto que seu intervalo está estimado entre 11 e 16%. Considerando, por fim, o Cenário 4, nota-se que no contexto de eventos, a taxa de falsos alarmes se concentra em valores menores, estimados entre 5% e 8% para o Chebyshev Adaptado 4% e 7% para Correlação. Para mais detalhes, a Tabela 6.3 apresenta a faixa de valores de acurácia média inferidos para cada cenário.

Nota-se inicialmente que a taxa de falsos alarmes para o contexto de sensores falhos, está relacionada à medida de similaridade em conjunto aos cenários existentes, principalmente para os cenários de falhas. Isso se percebe no fato de que enquanto existe uma variação maior entre cenários em relação ao Chebyshev Adaptado, na Correlação, em termos de nós falhos, os intervalos de confiança se cruzam.

Importante também ressaltar, mais uma vez, o contexto do Cenário 3, em que as anormalidades vão surgindo de maneira crescente entre os dados, o que acaba induzindo o sistema *fuzzy* a tomar decisões equivocadas, gerando alertas indevidos, pois os valores considerados

Cenario	**Média - FA**	**Limite Inferior**	**Limite Superior**	**Medida de Similaridade**
1	9.74	8.09	11.61	Chebyshev Adaptado (CA)
2	9.22	7.62	10.87	Chebyshev Adaptado (CA)
3	14.08	11.56	16.73	Chebyshev Adaptado (CA)
4	7.38	5.97	8.89	Chebyshev Adaptado (CA)
1	14.37	12.94	15.86	Correlação (CO)
2	14.79	13.17	16.44	Correlação (CO)
3	14.44	12.76	15.60	Correlação (CO)
4	6.19	4.99	7.48	Correlação (CO)

Tabela 6.3: Valores Estimados da Taxa de Falsos Alarmes

fora do padrão acabam se tornando normais ao sistema, trazendo erros na identificação.

Já para o cenário de eventos, nota-se que apesar de uma limitação no encontro de eventos, principalmente com a medida de Correlação, isso não se reflete na taxa de falsos alarmes. A dependência em relação às medidas de similaridade e as técnicas de detecção de anomalias, acabam nesse sentido não refletindo na geração de falsos alarmes, visto que um dos fatores que minam a efetividade no encontro de eventos está relacionado à técnica de detecção que muitas vezes falha em apontar a existência de anomalias, refletindo de maneira contrária a taxa de falsos alarmes. Além de que, por não existir o elemento *fuzzy* para esta tomada de decisão, os alertas produzidos com resultado de não-definido (como está descrito na seção 4.4.3) são praticamente inexistentes.

6.2 Quantidade de Sensores Falhos

Como já descrito nas seções 1, 2 e 3, uma das dificuldades dos mecanismos de categorização ou de detecção de nós falhos que se utilizam da informação dos dados vizinhança para tomar

decisões, está ligada a quando a quantidade de sensores falhos no *cluster* deixam de ser minoria, e se tornam metade ou maioria. Considerando esses ambientes, foi realizada uma análise pra verificar como o mecanismo proposto neste trabalho se comporta nestas três circunstâncias. Considerando, assim, três grupos: vizinhanças que tiveram menos de 50% dos sensores com falhas, vizinhanças que tiveram em torno de 50% dos seus sensores com falhas e vizinhanças que tiveram mais de 50% dos seus sensores com falhas. Os dados das métricas foram estimados com 90% de confiança, considerando os dados dos cenários já apresentados nas seções anteriores.

6.2.1 Acurácia

A Figura 6.4 apresenta os dados de acurácia estimados para os três tipos de circunstâncias possíveis. Com o medida de similaridade Chebyshev Adaptado o valor para as vizinhanças com menos da metade de sensores falhos ficou estimado entre 81% e 84%. No caso em que a vizinhança se divide entre nós sensores normais e nós sensores problemáticos, estima-se uma acurácia entre 71% e 80%. As duas estimativas apontam para a existência de diferença entre os grupos, que pode variar entre 1% a 13%. Considerando sensores falhos como maioria do *cluster*, estima-se valores de acurácia entre 62% e 66%. Podendo assim afirmar a existência de diferenças na média de acurácia para as três circunstâncias, com uma maior efetividade para o ambiente com sensores falhos minoritários, e uma efetividade menor para o ambiente em que os nós falhos são majoritários.

Ainda na Figura 6.4, considerando a utilização da correlação, para o grupo com menos de 50% de sensores falhos estima-se uma acurácia entre 76% e 80%, enquanto que para vizinhanças com 50% de nós falhos, os valores são estimados entre 70% e 77%, apresentando assim uma sobreposição nos intervalos de confiança, podendo desta forma haver diferença entre os dois grupos ou não. Por fim, estima-se entre 60% e 66% a acurácia para vizinhanças que possuem mais de 50% de sensores falhos em seu *cluster*, apresentando uma menor efetividade em comparação aos grupos anteriores.

No comparativo entre as medidas de similaridade, apenas para o contexto com sensores falhos minoritários é que se pode afirmar a existência de diferença, que pode variar de 1 a 8%, com menor efetividade para a utilização da Correlação. Nos demais contextos, os valores se interseccionaram.

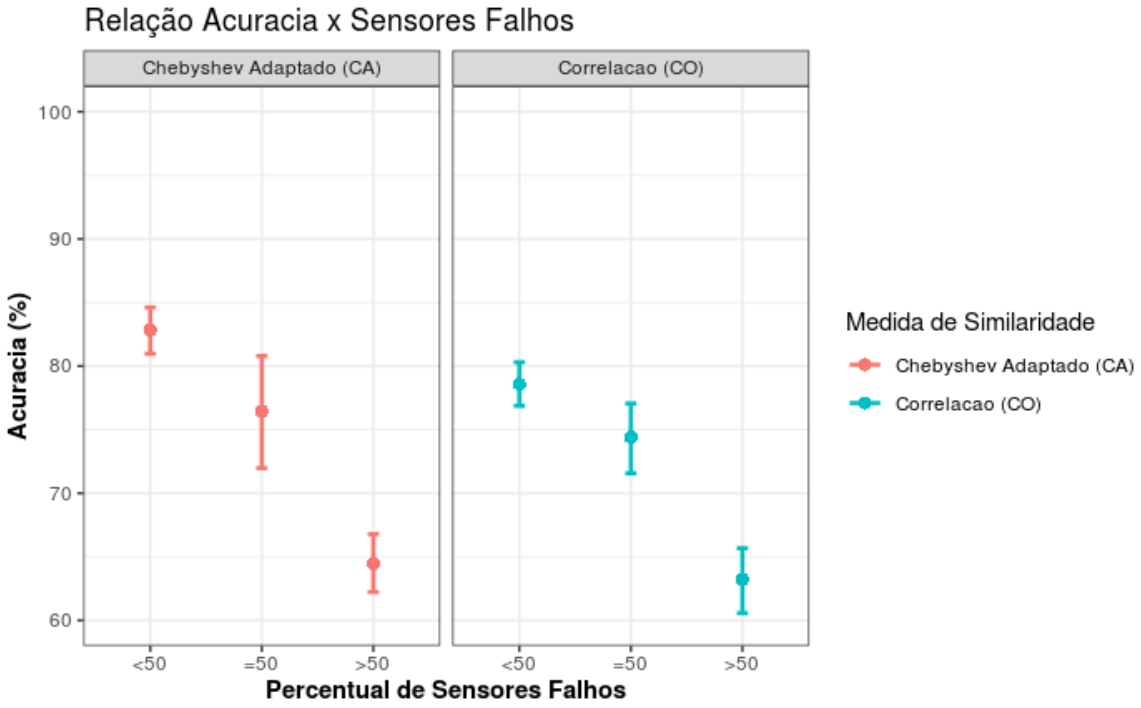

Figura 6.4: Acurácia versus Sensores Falhos.

6.2.2 Recall

Utilizando a medida de similaridade Chebyshev Adaptada, estima-se que para vizinhanças com menos da metade dos sensores falhos o valor de Recall está entre 62% e 70%, enquanto que para contextos com divisão entre sensores falhos e corretos, o intervalo se estabeleça entre 63% e 73%, não havendo assim a capacidade de afirmar a existência de diferença entre os dois contextos, no que diz respeito à sua capacidade de encontrar corretamente sensores falhos. Já para quando os sensores falhos se tornam maioria, os resultados apontam para um intervalo entre 44% e 51%, apresentando então uma menor efetividade em relação aos outros dois contextos. A Figura 6.5 apresenta o comparativo da média estimada do Recall nos três contextos.

Ainda na Figura 6.5, considerando a medida de Correlação, para o contexto em que menos da metade de nós na vizinhança são falhos, o valor de Recall é estimado entre 56% e 65%, ao passo que para clusters divididos entre falhos e normais, o valor do Recall da abordagem está estimado entre 52% e 62%. Não permitindo, assim, afirmar a existência ou não de diferença na capacidade de detectar corretamente nós falhos, entre os dois grupos. Para o contexto de maioria de sensores falhos na região, temos o valor de Recall estimado

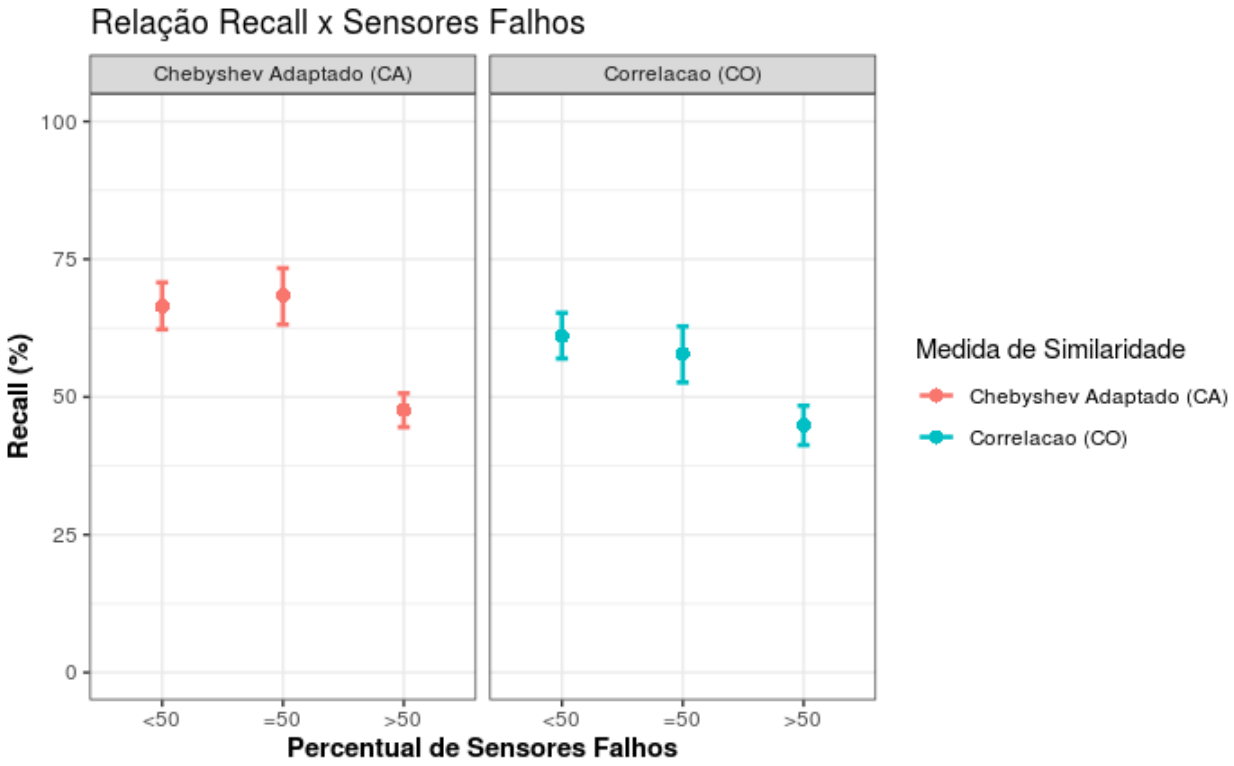

Figura 6.5: Recall versus Sensores Falhos.

entre 41% e 48%, apresentando assim uma distinção na eficiência se comparado aos grupos anteriores.

6.2.3 Taxa de Falsos Alarmes

A Figura 6.6 apresenta os valores estimados da taxa de falsos alarmes. Tendo em consideração a medida Chebyshev Adaptado, o valor estimado da taxa de falsos alarmes para *clusters* com minoria de sensores falhos se estabelece entre 11% e 15%. Já quando os sensores falhos são majoritários a taxa de falsos alarmes é estimada entre 13% e 17%. Desta forma, não se pode afirmar a existência de diferença entre os dois grupos. Já quando a vizinhança é dividida entre falhos e corretos, o valor estimado de falsos alarmes se estabelece entre 15% e 25%, também não sendo possível afirmar a existência ou não de diferença entre os grupos.

Considerando a utilização da Correlação como métrica de similaridade, não se pode afirmar a existência de diferença ou não entre os três grupos, no qual ao grupo com menos da metade de sensores falhos, o valor médio de falsos alarmes está estimado entre 17% e 20%, enquanto que para o grupo com 50% de sensores falhos o valor é presumido entre 17% e

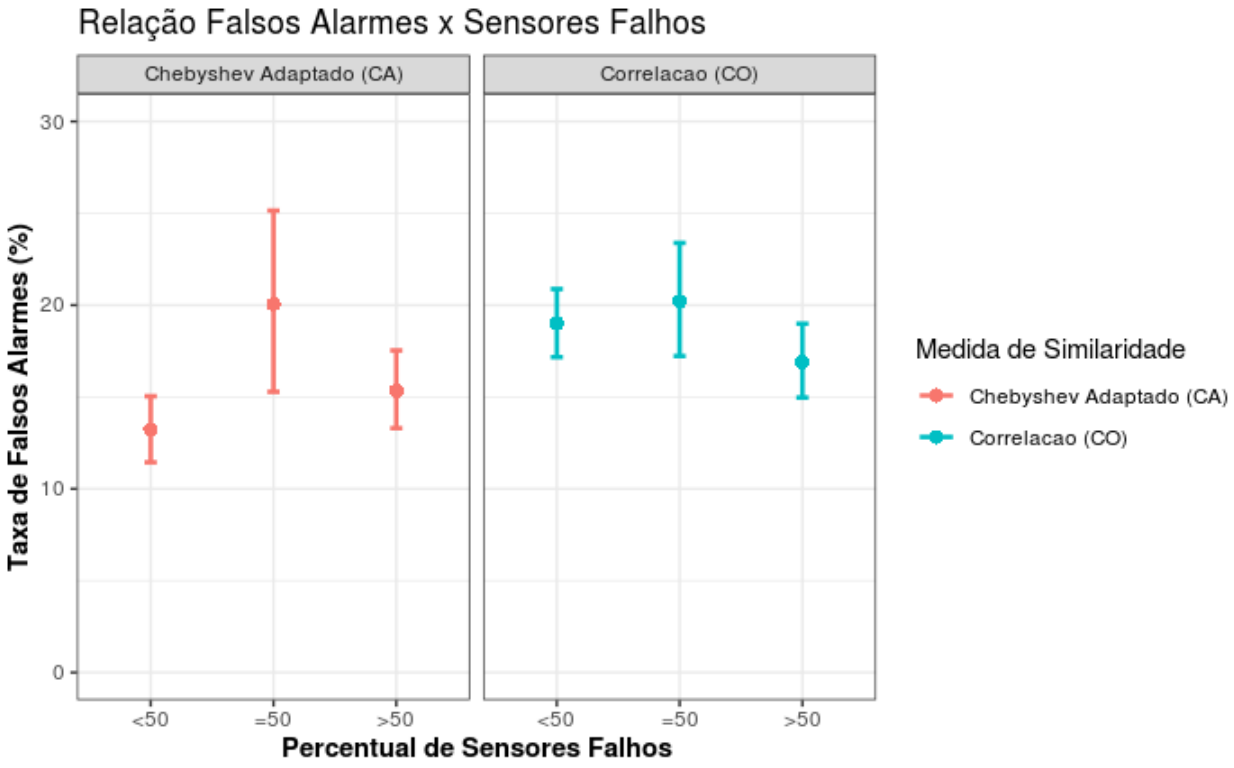

Figura 6.6: Falsos Alarmes versus Sensores Falhos.

23%, e para o grupo com sensores majoritariamente falhos no cluster, a taxa de alarme está contida entre 14% e 19%.

6.3 Perda de Pacotes

A fim de responder ainda as questões levantadas no Capítulo 5, também foi analisada a relação entre a eficiência da abordagem e a perda de pacotes nos *clusters*. Para isso, procurou-se analisar a distribuição e a correlação (utilizando a correlação de spearman) entre os resultados das métricas das vizinhanças e suas respectivas replicações, que foram apresentados anteriormente, e a taxa de perda de pacotes, que é a razão entre a quantidade de pacotes recebidos na estação base ou no *clusterhead* (para Eventos) e a quantidade de pacotes enviados.

6.3.1 Acurácia

A Figura 6.7 apresenta a distribuição dos dados médios de perda de pacotes das vizinhanças em relação aos valores médios de acurácia. A distribuição nos mostra que existe uma corre-

lação moderada entre a taxa de perda de pacote e a acurácia, utilizando a medida Chebyshev Adaptado. Aplicando a correlação de Spearman temos uma correlação negativa moderada de -0,43. Já aplicando a medida de correlação na nossa abordagem, também se percebe um comportamento correlacionado no limiar entre moderado e fraco, menor evidência de relação de que a medida anterior, com valor de -0,39.

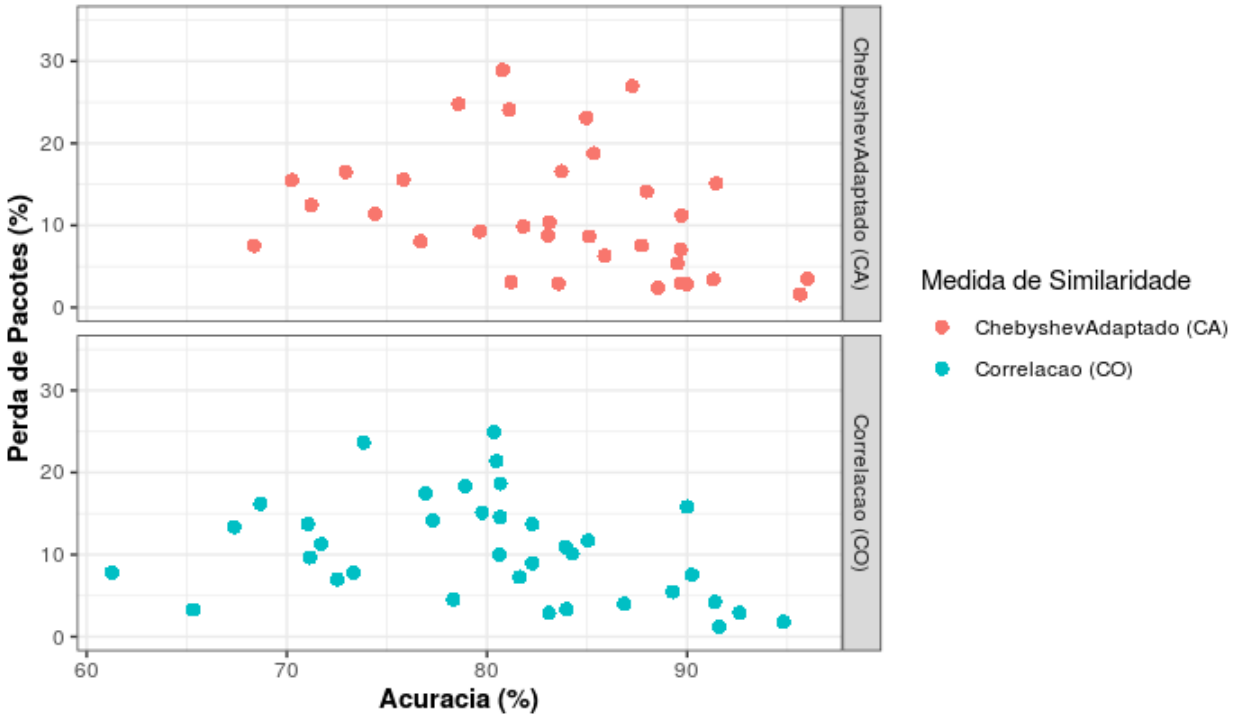

Figura 6.7: Acurácia versus Perda de Pacotes

.

6.3.2 Recall

Os dados de Recall, se comportam de maneiras um pouco distintas considerando as medidas de similaridade utilizadas. Considerando a medida de similaridade CA, se percebe uma relação entre moderada e fraca entre a perda de pacotes e a capacidade de encontrar corretamente nós falhos e eventos pela abordagem. Aplicando a correlação de Spearman, para conseguirmos mensurar de maneira mais palpável tal relação, temos como resultado uma correlação negativa de -0.39, que se estabelece no limiar entre fraco e moderado. Considerando a medida de similaridade CO, é possível afirmar que a relação entre o Recall e a perda de pacotes

é fraca, apresentando o índice de Spearman de -0.25. Desta maneira, não é possível aferir com certeza uma alta influencia da perda de pacotes nos resultados de Recall.

A Figura 6.8 apresenta a distribuição dos dados da relação Recall versus perda de pacotes.

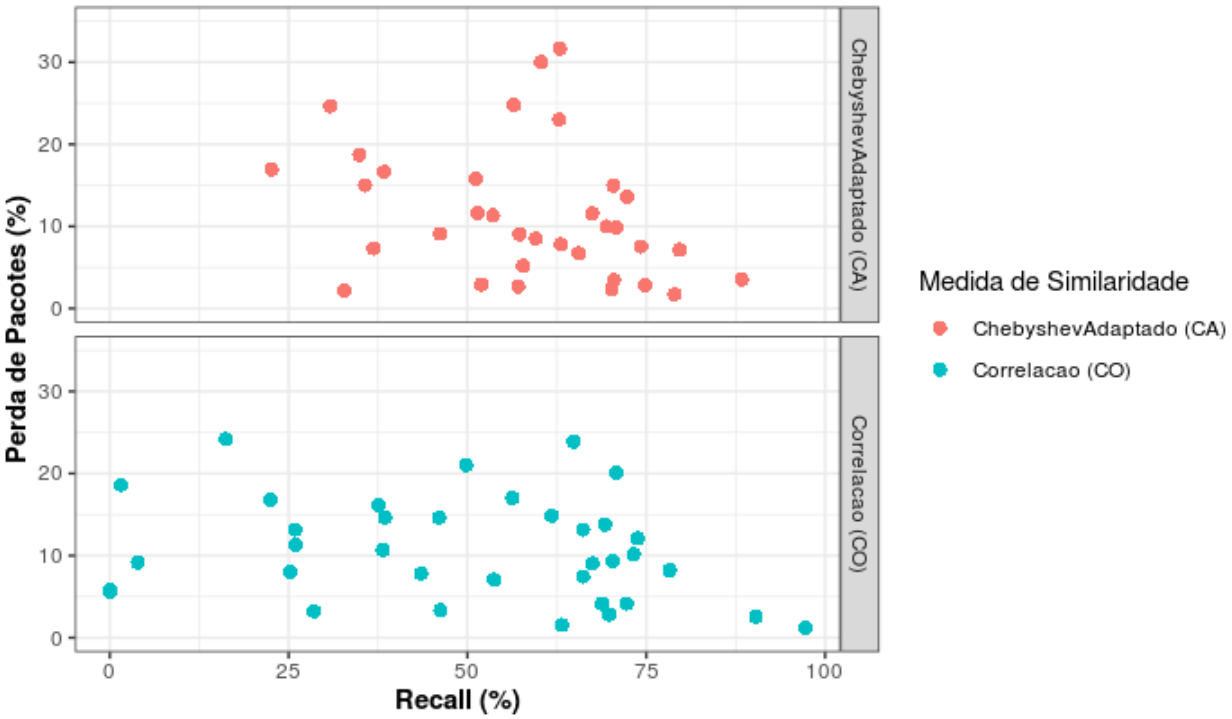

Figura 6.8: Recall versus Perda de Pacotes.

6.3.3 Taxa de Falsos Alarmes

Considerando a Taxa de Falsos alarmes, é perceptível uma relação moderada entre as variáveis, como mostra a Figura 6.9. Para o Chebyshev adaptado, o coeficiente de Spearman aponta para uma correlação moderada positiva de 0.50, tendo uma atenuação nesse comportamento pelo maiores valores médios de falsos alarmes, que não estão ligados aos maiores valores de perda de pacotes.

Quando foi utilizado a Correlação como métrica de similaridade, o coeficiente se apresenta com 0.43, indicando um certo nível de influência entre a perda de pacotes com a geração de falsos positivos, entretanto não sendo um fator totalmente relevante para definição dos resultados.

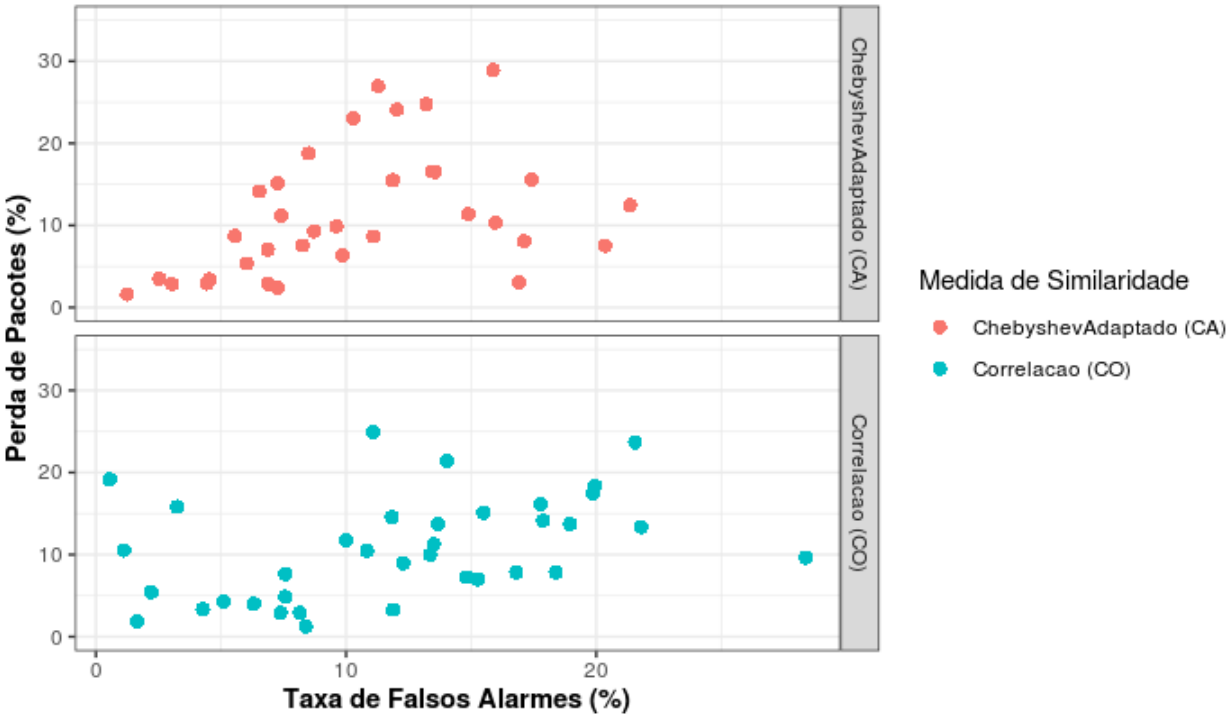

Figura 6.9: Falsos Alarmes versus Perda de Pacotes.

6.4 Discussão

Esta seção resume os achados referentes às respostas das questões de pesquisa, elaboradas no Capítulo 5 e relembradas aqui:

- **Questão de pesquisa 1 - QP1**: A abordagem consegue categorizar as anomalias de maneira eficiente?
- **Questão de Pesquisa 2 - QP2**: No contexto de falhas nos dados, a abordagem consegue categorizar falhas de diferentes formatos nos sensores com problemas?
- **Questão de Pesquisa 3 - QP3**: A abordagem consegue determinar a existência de sensores falhos à medida que em que eles vão se tornando maioria na vizinhança?
- **Questão de Pesquisa 4 - QP4:** A abordagem sofre algum tipo de atenuação devido a perda de dados ocorrida na rede?

QP1. Considerando todos os aspectos levantados, é possível afirmar que nos cenários avaliados a abordagem consegue agir de maneira eficaz na categorização das anomalias. Levando em conta a acurácia, todos os contextos tiveram uma estimativa média de pelo menos 69% na taxa de acertos. Já olhando aspectos relacionados a alertar efetivamente a existência de eventos ou sensores falhos, percebeu-se uma maior eficácia para contextos falhos intermitentes do que em relação a eventos. Especificamente em eventos, verifica-se também uma forte relação e dependência na categorização efetiva dos eventos com as medidas de similaridade utilizadas, como foi exposto os resultados com menores níveis e com intervalos mais amplos quando a medida de similaridade utilizada foi a Correlação (CO).

QP2. Dados os resultados obtidos, é possível afirmar que a abordagem consegue categorizar os sensores falhos nos três cenários, entretanto, através dos valores observados de Recall e Taxa de falsos alarmes, é também possível afirmar que esta efetividade é diminuída para o Cenário 3, em que as anomalias vão se tornando crescentes à medida em que o tempo passa. Para os cenários com falhas intermitentes, nota-se uma efetividade parecida, e que não existe influência entre essas anomalias serem mais ou menos frequentes.

QP3. A partir das métricas de análise, percebe-se que em todos os contextos relativos à quantidade de sensores falhos presentes, houve efetiva categorização de anomalias. Entretanto, é necessário considerar que quando os sensores falhos são maioria, ou seja, são mais do que 50% dos sensores do cluster, a eficácia da abordagem na categorização é reduzida. Contudo, em casos de divisão entre sensores, ou seja, metade de sensores falhos e metade de sensores normais, não é possível afirmar diferença entre este cenário com o cenário em que nós problemáticos são minoritários.

QP4. Os resultados apontam que, considerando a geração de falsos positivos, existe uma relação moderada entre a perda de pacotes e a geração de falsos alarmes, assim como para acurácia geral. Entretanto, essa relação é vista em relação ao Recall apenas para a utilização do Chebyshev Adaptado como medida de similaridade, que se estabelece no limiar entre fraco e moderado. Enquanto que a relação é fraca, para a medida de similaridade Correlação, não sendo possível assim afirmar a relação entre perda de pacotes e o encontro correto de sensores falhos ou eventos. De forma geral, algumas considerações podem ser feitas com os resultados apresentados. Considerando os cenários relacionados a falhas, nota-se que a

abordagem apresenta uma menor efetividade para Cenário 3. Isso se dá pelo fato de que quando os dados nas janelas se tornam apenas anômalos, o sistema considerará o padrão anômalo como normal, prejudicando tanto o modelo preditor quanto a técnica de detecção de anomalias. O que levanta o ponto da necessidade de medidas de atuação para evitar que a medida em que alertas de sensores falhos sejam lançados, ajustes ou isolamento de sensores sejam feitos impedindo que os dados anormais não se tornem o padrão da rede, enviesando assim toda possível análise do que está ocorrendo.

Considerando o cenário de eventos, nota-se que existe uma dependência maior para este cenário com as medidas e técnicas utilizadas. Sem o elemento *fuzzy*, invariavelmente a detecção dos eventos existentes fica relacionada a como a técnica de detecção de anomalias vai identificar a existências dos *outliers* e de como as medidas de similaridades vão comparar os sensores. Com isso percebe-se que medidas de similaridades mais sensíveis a oscilações nos dados, como é o caso da Correlação, que analisa tendências, faz com que a identificação de eventos se reduza e se torne mais variável. Permitindo assim a análise de ajustes que podem ser feitos, a incluir elementos e informações sobre eventos no sistema *fuzzy* criado na segunda etapa da abordagem, atenuando assim, a dependência desses casos em relação as técnicas de detecção e as medidas de similaridade empregadas.

Capítulo 7

Considerações Finais e Trabalhos Futuros

Este trabalho foi desenvolvido com o objetivo de propor alternativas para a categorização online de anomalias no ambiente de redes de sensores sem fio. Para tal, foi desenvolvido um mecanismo hibrido/distribuído baseado em um sistema de lógica *fuzzy* e detecção espaço/-temporal de anomalias, a partir de características ligadas à identificação de anormalidades, similaridade de sensores e histórico dos dados. O mecanismo apresentado tem por intuito propor um meio de identificar a existência de eventos e sensores falhos, inclusive em contextos onde os sensores falhos são maioria em um cluster.

A abordagem foi construída em dois níveis. O primeiro consegue atestar a ocorrência de eventos ou não, baseado em detecções de anomalias ocorridas em cada sensor de um *cluster* e de comparação entre os dados desses sensores, podendo assim dado a existência de anormalidades ou não, e da similaridade ou não definir se existe um evento naquele momento.

Devido às limitações existentes para definir sensores falhos apenas pelas comparações entre os dados dos sensores vizinhos, visto que muitos sensores falhos podem surgir no ambiente e enviesar a comparação e a detecção, uma segunda camada é construída baseada em lógica *fuzzy*. Nesta segunda etapa, um algoritmo de predição/classificação é utilizado para aferir a relação entre janela de dados atual, com seu histórico. Dado isto, o sistema fuzzy constrói sua inferência a partir de valores de similaridade entre sensores, relação entre histórico e conjunto de dados atual e a relação dos dados antecedentes dos vizinhos diferentes com suas respectivas janelas de dados. Permitindo, assim, a identificação de sensores falhos

ou não.

A fim de avaliar a solução mencionada, um conjunto de experimentos foi realizado, considerando uma simulação de aplicação em uma rede de sensores sem fio industrial, a fim de avaliar cenários distintos na existência de falhas em sensores e eventos. A abordagem, então, demonstrou efetividade no encontro e na categorização das anomalias no ambiente de redes de sensores sem fio online, com valores médios estimados de acurácia acima de 69% para todos os cenários. Com atenuações nesta efetividade para contextos específicos ligados ao tipo de medida de similaridade escolhida e o cenário, como anomalias crescentes ao longo do tempo e eventos.

7.1 Limitações

Considerando que o escopo definido no presente trabalho resultou em determinadas limitações, é notório que o mecanismo proposto tem potencial e capacidade de melhorias a fim de obter melhores resultados. Desta maneira, com o intuito de explorar ainda mais os problemas supracitados no trabalho, expandindo novos conceitos com base nos resultados apresentados e incrementando novos fatores com novas pesquisas, as principais limitações são elencadas a seguir:

- O mecanismo foi testado apenas utilizando um tipo de algoritmo de detecção de anomalias e um tipo de mecanismo preditor, sendo necessário avaliações com outros tipos de algoritmos para potenciais evoluções de resultados e confirmações da influência desses algoritmos nos resultados expostos;

- Não foi analisado que tipos de medidas podem ser tomadas para mitigar e solucionar problemas referentes aos sensores falhos dentro do próprio sistema, a fim de atuar e gerar a manutenção dos sensores considerados falhos de maneira automática;

- O trabalho não avaliou a relação de medidas do sistema *fuzzy* em relação aos eventos, e que tipos de regras podem ser incluídas a fim de existir mais um elemento de determinação e confirmação da existência de eventos ou não;

- Não foram analisadas questões relacionada a custos de processamento dos algoritmos

nos nós sensores, sendo necessário a investigação deste custo para eventuais implementações reais futuras.

7.2 Trabalhos Futuros

Conforme apresentado na Seção 7.1, alguns elementos podem ajudar na evolução deste trabalho. Com isso, pode-se destacar alguns trabalhos futuros importantes:

- Avaliar o comportamento da abordagem, com uma maior variedade de algoritmos, técnicas e medidas dentro dos elementos da abordagem, seja na detecção, na determinação de similaridade ou na predição e, assim, estabelecer com maior confiança o grau de influência desses algoritmos na melhora ou piora da categorização das anomalias;
- Avaliar o custo computacional dos algoritmos;
- Avaliar que tipos de novos elementos podem ser integrados ao sistema de inferência fuzzy, para que o mesmo também seja utilizado na identificação de eventos;
- Construir mecanismos automatizados de manutenção e atuação na rede, que mitiguem as limitações em cenários em que as anomalias ao longo do tempo se tornem o padrão dos dados, permitindo assim uma maior efetividade da abordagem.

Bibliografia

[1] Wireless sensors network market - growth, trends, covid-19 impact, and forecasts (2021 - 2026), year = 2020,

[2] E. Aleskerov, B. Freisleben, and B. Rao. Cardwatch: a neural network based database mining system for credit card fraud detection. In *Proceedings of the IEEE/IAFE 1997 Computational Intelligence for Financial Engineering (CIFEr)*, pages 220–226, March 1997.

[3] Mennatallah Amer. Comparison of unsupervised anomaly detection techniques bachelor thesis. 2011.

[4] Giuseppe Anastasi, Marco Conti, Mario Di Francesco, and Andrea Passarella. Energy conservation in wireless sensor networks: A survey. *Ad Hoc Networks*, 7(3):537 – 568, 2009.

[5] Patrícia Bordignon André et al. Detecção e identificação de outliers em redes de sensores sem fio de larga escala. 2017.

[6] Linnyer Beatrys Ruiz Raquel Aparecida de Freitas Mini Eduardo Freire Nakamura Carlos Maurıcio Serodio Figueiredo Antonio A.F. Loureiro, José Marcos S. Nogueira. Redes de sensores sem fio. *XXI Simpósio Brasileiro de Redes de Computadores*, 21:179–226, 2003.

[7] Majid Bahrepour, Nirvana Meratnia, Mannes Poel, Zahra Taghikhaki, and Paul JM Havinga. Distributed event detection in wireless sensor networks for disaster management. In *2010 international conference on intelligent networking and collaborative systems*, pages 507–512. IEEE, 2010.

[8] A. Boulis. Castalia a simulator for wireless sensor networks and body area networks - user's manual version 3.0. 2010.

[9] Shihua Cao, Qihui Wang, Yaping Yuan, and Junyang Yu. Anomaly event detection method based on compressive sensing and iteration in wireless sensor networks. *Journal of Networks*, 9(3):711, 2014.

[10] Prasenjit Chanak and Indrajit Banerjee. Fuzzy rule-based faulty node classification and management scheme for large scale wireless sensor networks. *Expert Systems with Applications*, 45:307–321, 2016.

[11] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Comput. Surv.*, 41(3):15:1–15:58, July 2009.

[12] Jinran Chen, Shubha Kher, and Arun Somani. Distributed fault detection of wireless sensor networks. In *Proceedings of the 2006 workshop on Dependability issues in wireless ad hoc networks and sensor networks*, pages 65–72, 2006.

[13] Hongju Cheng, Danyang Feng, Xiaobin Shi, and Chongcheng Chen. Data quality analysis and cleaning strategy for wireless sensor networks. *EURASIP Journal on Wireless Communications and Networking*, 2018(1):1–11, 2018.

[14] Ruan Delgado Gomes, Marcéu Oliveira Adissi, Abel Cavalcante Lima-Filho, Marco Aurélio Spohn, and Francisco Antônio Belo. On the impact of local processing for motor monitoring systems in industrial environments using wireless sensor networks. *International Journal of Distributed Sensor Networks*, 9(7):471917, 2013.

[15] Milan Erdelj, Nathalie Mitton, Enrico Natalizio, et al. Applications of industrial wireless sensor networks. *Industrial wireless sensor networks: applications, protocols, and standards*, pages 1–22, 2013.

[16] Asmaa Fawzy, Hoda MO Mokhtar, and Osman Hegazy. Outliers detection and classification in wireless sensor networks. *Egyptian Informatics Journal*, 14(2):157–164, 2013.

[17] Ruan Gomes, Diego Queiroz, Iguatemi Fonseca, and Marcelo Alencar. A simulation model for industrial multi-channel wireless sensor networks. *Journal of Communication and Information Systems*, 32(1), May 2017.

[18] Lin Gu, Dong Jia, Pascal Vicaire, Ting Yan, Liqian Luo, Ajay Tirumala, Qing Cao, Tian He, John A Stankovic, Tarek Abdelzaher, et al. Lightweight detection and classification for wireless sensor networks in realistic environments. In *Proceedings of the 3rd international conference on Embedded networked sensor systems*, pages 205–217, 2005.

[19] Vehbi C Gungor and Gerhard P Hancke. Industrial wireless sensor networks: Challenges, design principles, and technical approaches. *IEEE Transactions on industrial electronics*, 56(10):4258–4265, 2009.

[20] Hoaglin D. Iglewicz, B. *How to detect and handle outliers*. ASQC Quality Press, 1993.

[21] Gonçalo Jesus, António Casimiro, and Anabela Oliveira. A survey on data quality for dependable monitoring in wireless sensor networks. *Sensors*, 17(9):2010, 2017.

[22] Sai Ji, Shen-fang Yuan, Ting-huai Ma, and Chang Tan. Distributed fault detection for wireless sensor based on weighted average. In *2010 Second International Conference on Networks Security, Wireless Communications and Trusted Computing*, volume 1, pages 57–60. IEEE, 2010.

[23] P Johnson and D C Andrews. Remote continuous physiological monitoring in the home. *Journal of Telemedicine and Telecare*, 2(2):107–113, 1996.

[24] Raja Jurdak, X. Rosalind Wang, Oliver Obst, and Philip Valencia. *Wireless Sensor Network Anomalies: Diagnosis and Detection Strategies*, pages 309–325. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[25] Aimad Karkouch, Hajar Mousannif, Hassan Al Moatassime, and Thomas Noel. Data quality in internet of things. *Journal of Network and Computer Applications*, 73:57–81, 2016.

[26] Ehsan Khazaei, Ali Barati, and Ali Movaghar. Improvement of fault detection in wireless sensor networks. In *2009 ISECS International Colloquium on Computing, Communication, Control, and Management*, volume 4, pages 644–646, 2009.

[27] S. Krco, M. Johansson, V. Tsiatsis, I. Cubic, K. Matusikova, and R. Glitho. Mobile network supported wireless sensor network services. In *2007 IEEE International Conference on Mobile Adhoc and Sensor Systems*, pages 1–3, Oct 2007.

[28] V. Kumar. Parallel and distributed computing for cybersecurity. *IEEE Distributed Systems Online*, 6(10), 2005.

[29] Edgar Noschang Kunz et al. Análise de séries temporais: estudo estatístico sobre modelos arima com uma aplicação prática em processo sazonal determinístico. 2020.

[30] S. Lan, M. Qilong, and J. Du. Architecture of wireless sensor networks for environmental monitoring. In *2008 International Workshop on Education Technology and Training 2008 International Workshop on Geoscience and Remote Sensing*, volume 1, pages 579–582, Dec 2008.

[31] R.S. LANZILLOTTI. *Lógica Fuzzy: uma Abordagem Para Reconhecimento de Padrão*. Paco Editorial, 2014.

[32] Leandro da Costa Moraes Leite et al. Geração e simplificação da base de conhecimento de um sistema híbrido fuzzy-genético. 2009.

[33] Mo Li, Yunhao Liu, and Lei Chen. Nonthreshold-based event detection for 3d environment monitoring in sensor networks. *IEEE Transactions on Knowledge and Data Engineering*, 20(12):1699–1711, 2008.

[34] Qilian Liang and Lingming Wang. Event detection in sensor networks using fuzzy logic system. In *EEE International Conference on Computational Intelligence for Homeland Security and Personal Safety, Orlando, FL, USA*, 2005.

[35] Bin Lu and Vehbi C Gungor. Online and remote motor energy monitoring and fault diagnostics using wireless sensor networks. *IEEE Transactions on Industrial Electronics*, 56(11):4651–4659, 2009.

[36] Samuel Madden, Michael J. Franklin, Joseph M. Hellerstein, and Wei Hong. The design of an acquisitional query processor for sensor networks. In *Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data*, SIGMOD '03, pages 491–502, New York, NY, USA, 2003. ACM.

[37] Abinash Mahapatra, Kumar Anand, and Dharma P. Agrawal. Qos and energy aware routing for real-time traffic in wireless sensor networks. *Computer Communications*, 29(4):437 – 445, 2006. Current areas of interest in wireless sensor networks designs.

[38] Gary M Marsh and Songwon Seo. A review and comparison of methods for detecting outliers in univariate data sets. 2006.

[39] Joseph E. Mbowe and George S. Oreku. Quality of service in wireless sensor networks. 2014.

[40] C. O'Reilly, A. Gluhak, M. A. Imran, and S. Rajasegarar. Anomaly detection in wireless sensor networks in a non-stationary environment. *IEEE Communications Surveys Tutorials*, 16(3):1413–1432, 2014.

[41] Colin O'Reilly, Alexander Gluhak, Muhammad Ali Imran, and Sutharshan Rajasegarar. Anomaly detection in wireless sensor networks in a non-stationary environment. *IEEE Communications Surveys & Tutorials*, 16(3):1413–1432, 2014.

[42] M. Perillo and W. B. Heinzelman. Providing application qos through intelligent sensor management. In *Proceedings of the First IEEE International Workshop on Sensor Network Protocols and Applications, 2003.*, pages 93–101, May 2003.

[43] V. Potdar, A. Sharif, and E. Chang. Wireless sensor networks: A survey. In *2009 International Conference on Advanced Information Networking and Applications Workshops*, pages 636–641, May 2009.

[44] D. Puccinelli and M. Haenggi. Wireless sensor networks: applications and challenges of ubiquitous sensing. *IEEE Circuits and Systems Magazine*, 5(3):19–31, 2005.

[45] Diego V Queiroz, Marcelo S Alencar, Ruan D Gomes, Iguatemi E Fonseca, and Cesar Benavente-Peces. Survey and systematic mapping of industrial wireless sensor networks. *Journal of Network and Computer Applications*, 97:96–125, 2017.

[46] Diego Véras de Queiroz et al. Simulação realista de redes de sensores sem fio industriais. 2016.

[47] R. Rajagopalan and P. K. Varshney. Data-aggregation techniques in sensor networks: A survey. *IEEE Communications Surveys Tutorials*, 8(4):48–63, Fourth 2006.

[48] S. Rajasegarar, C. Leckie, and M. Palaniswami. Anomaly detection in wireless sensor networks. *IEEE Wireless Communications*, 15(4):34–40, 2008.

[49] S. Rajasegarar, C. Leckie, M. Palaniswami, and J. C. Bezdek. Quarter sphere based distributed anomaly detection in wireless sensor networks. *IEEE International Conference on Communications*, pages 3864–3869, 2007.

[50] Priyanka Rawat, Kamal Deep Singh, Hakima Chaouchi, and Jean-Marie Bonnin. Wireless sensor networks: a survey on recent developments and potential synergies. *The Journal of Supercomputing*, 68:1–48, 2013.

[51] Biljana Risteska Stojkoska, Dimitar Solev, and Danco Davcev. Data prediction in wsn using variable step size lms algorithm. 08 2011.

[52] Hesam Sagha, Jose del R Mill, Ricardo Chavarriaga, et al. Detecting and rectifying anomalies in body sensor networks. In *2011 International Conference on Body Sensor Networks*, pages 162–167. IEEE, 2011.

[53] Bruno P. Santos, Lucas A. M. Silva, Clayson S. F. S. Celes, Joao B. Borges Neto, Bruna S. Peres, Marcos Augusto M. Vieira, Luiz Filipe M. Vieira, Olga N. Goussevskaia, and Antonio A.F. Loureiro. Internet das coisas: da teoria à prática. *Em Minicursos do XXXIV SBRC*, pages 1–50, 2016.

[54] SBC. Grandes desafios da pesquisa em computação no brasil – 2006-2016, 2006.

[55] Umrah Shafeeq and Prasenjit Chanak. Heterogeneous hardware fault-detection scheme for large scale wireless sensor networks. In *2018 International Conference on Circuits and Systems in Digital Enterprise Technology (ICCSDET)*, pages 1–5, 2018.

[56] Nauman Shahid, Ijaz Haider Naqvi, and Saad Bin Qaisar. Svm based event detection and identification: exploiting temporal attribute correlations using sensgru. *Mathematical Problems in Engineering*, 2014, 2014.

[57] S. Sharma, R. K. Bansal, and S. Bansal. Issues and challenges in wireless sensor networks. In *2013 International Conference on Machine Intelligence and Research Advancement*, pages 58–62, Dec 2013.

[58] Yashwant Singh, Suman Saha, Urvashi Chugh, and Chhavi Gupta. Distributed event detection in wireless sensor networks for forest fires. In *2013 UKSim 15th International Conference on Computer Modelling and Simulation*, pages 634–639. IEEE, 2013.

[59] C. Spence, L. Parra, and P. Sajda. Detection, synthesis and compression in mammographic image analysis with a hierarchical image probability model. In *Proceedings IEEE Workshop on Mathematical Methods in Biomedical Image Analysis (MMBIA 2001)*, pages 3–10, Dec 2001.

[60] Ryo Sugihara and Rajesh K. Gupta. Programming models for sensor networks: A survey. *ACM Trans. Sen. Netw.*, 4(2):8:1–8:29, April 2008.

[61] Pedro Lemos Tavares. Redes de sensores sem-fio. https://www.gta.ufrj.br/grad/02_2/Redesdesensores/RedesdeSensoresSem-fio.html,, 2002.

[62] Valiana Alves Teodoro. *Modelos de séries temporais para temperatura em painéis de cimento-madeira*. PhD thesis, Universidade de São Paulo, 2015.

[63] Chinh T Vu, Raheem A Beyah, and Yingshu Li. Composite event detection in wireless sensor networks. In *2007 IEEE International Performance, Computing, and Communications Conference*, pages 264–271. IEEE, 2007.

[64] Geoffrey Werner-Allen, Konrad Lorincz, Mario Ruiz, Omar Marcillo, Jeff Johnson, Jonathan Lees, and Matt Welsh. Deploying a wireless sensor network on an active volcano. *IEEE internet computing*, 10(2):18–25, 2006.

[65] Felix Wortmann and Kristina Flüchter. Internet of things. *Business & Information Systems Engineering*, 57(3):221–224, 2015.

[66] Miao Xie, Song Han, Biming Tian, and Sazia Parvin. Anomaly detection in wireless sensor networks: A survey. *Journal of Network and Computer Applications*, 34(4):1302 – 1325, 2011.

[67] Vivian Mayumi Yamassaki. Combinando modelos de machine learning com lógica fuzzy, nov 2020.

[68] J. Yick, B. Mukherjee, and D. Ghosal. Wireless sensor network survey. *The International Journal of Computer and Telecommunications Networking*, 52(12):2292–2330, 2008.

[69] Hao Yuan, Xiaoxia Zhao, and Liyang Yu. A distributed bayesian algorithm for data fault detection in wireless sensor networks. In *2015 International Conference on Information Networking (ICOIN)*, pages 63–68, 2015.

[70] J. Zhang, W. Li, Z. Yin, S. Liu, and X. Guo. Forest fire detection system based on wireless sensor network. In *2009 4th IEEE Conference on Industrial Electronics and Applications*, pages 520–523, May 2009.

Apêndice A

Códigos Fonte - Castalia

Para o acesso dos códigos utilizados em toda experimentação no ambiente do Simulador Castalia, assim como os arquivos com os resultados acessar o repositório no github: https://github.com/MiqueasGaldino/AbordagemFuzzy