Aujor Tadeu Cavalca Andrade

Uma Abordagem Orientada a Dados para Reconfiguração de Topologia de Rede Cluster-Tree

Florianópolis (SC) - Brasil

2018

Aujor Tadeu Cavalca Andrade

Uma Abordagem Orientada a Dados para Reconfiguração de Topologia de Rede Cluster-Tree

Tese submetida ao Programa de Pós-Graduação em Engenharia de Automação e Sistemas da Universidade Federal de Santa Catarina para a obtenção do Grau de Doutor em Engenharia de Automação de Sistemas.

Universidade Federal de Santa Catarina – UFSC

Departamento de Automação e Sistemas

Programa de Pós-Graduação em Engenharia de Automação e Sistemas

Orientador: Prof. Dr. Carlos Barros Montez

Coorientador: Prof. Dr. Ricardo Alexandre Reinaldo de Moraes

Florianópolis (SC) - Brasil

2018

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Cavalca Andrade, Aujor Tadeu
Uma Abordagem Orientada a Dados para Reconfiguração de Topologia de Redes Cluster-Tree / Aujor Tadeu Cavalca Andrade ; orientador, Carlos Barros Montez, coorientador, Ricardo Alexandre Reinaldo de Moraes, 2019.
218 p.

Tese (doutorado) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós Graduação em Engenharia de Automação e Sistemas, Florianópolis, 2019.

Inclui referências.

1. Engenharia de Automação e Sistemas. 2. Detecção de Eventos. 3. Detecção de Outlier. 4. Estratégia de monitoramento. 5. Cluster-tree baseada em dados. I. Barros Montez, Carlos . II. de Moraes, Ricardo Alexandre Reinaldo . III. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Engenharia de Automação e Sistemas. IV. Título.

Aujor Tadeu Cavalca Andrade

Uma Abordagem Orientada a Dados para Reconfiguração de Topologia de Rede Cluster-Tree

Tese submetida ao Programa de Pós-Graduação em Engenharia de Automação e Sistemas da Universidade Federal de Santa Catarina para a obtenção do Grau de Doutor em Engenharia de Automação de Sistemas.

Florianópolis (SC), 20 de Março de 2019.

Banca Examinadora:

Prof. Dr. Werner Krauss Juniorr
Coordenador do Programa de Pós-Graduação em Engenharia de Automação e Sistemas

---

Prof. Dr. Carlos Barros Montez
Orientador

Prof. Dr. Ricardo Alexandre Reinaldo de Moraes
Co-orientador

Prof. Dr. Luiz Affonso H. Guedes
Convidado (Videoconferência)

Prof. Dr. Marcelo Ricardo Stemmer
Convidado 2

Prof. Dr. Gustavo Medeiros de Araújo
Convidado 3

Dedico esta tese de doutorado aos meus pais Aujor e Ivani,
e a meus irmãos Eduardo e Melissa.
Sem vocês não seria possível.

Agradecimentos

Primeiramente agradeço a Deus e aos meus pais, Aujor e Ivani pelo incentivo, valores e amor. Sou orgulho dos meus pais e dos ensinamentos que me permitiram alcançar essa conquista. Este trabalho é dedicado a vocês em agradecimento !!

Aos meus irmãos Eduardo e Melissa, pelo carinho e amor em cada momento desde pequeno até a vida adulta !! Tenho orgulho de vocês e gratidão por serem meus irmãos ! Este trabalho é dedicado a vocês em agradecimento !!

A minha Gabriela pelo amor, ouvidos e parceria ao final desta jornada. Foram muitas conversas e suas palavras sempre foram de incentivo, acolhimento e carinho. Este trabalho é dedicado a você em agradecimento!!

Aos professores Carlos Montez e Ricardo Moraes pela orientação, ensinamentos e parceria ao longo do doutorado, meu muito obrigado. Ao professor e colega Erico Leão pela ajudar parceira, conversar, sugestões ao longo do trabalho. Aos amigos e colegas Cristian e Angelo pelas conversar, ideias e parcerias.

Aos amigos de doutorado, Leonardo, André, Eduardo, Renan, Luiz, Regiane, Karilla, Patrícia, Miguel, Toscano, Sid, Flavio, Alexandre e todos os que participaram deste processo dividindo ideias e momentos.

Por fim aos professores PGEAS pelos ensinamento e ao Enio pela auxilio nos trâmites acadêmicos.

A persistência é o caminho do êxito.
Charles Chaplin.

Abstract

The Wireless Sensors Networks (WSNs) arise as an important research field, incorporating some related fields as communication areas, sensing, and data processing. The WSNs utilization scenario is wide, including environmental, industrial, and hospital monitoring. In this context, this thesis application scenario is wide areas monitoring, which includes challenging specificities as large spatial cover area, non-fixed sensors quantity per area, hardware implementation difficulty, and the possibility of generating a large number of outliers. The computational resources are limited at the nodes and associated with the potential hostile environments hardware installation and the environmental dynamic change in the monitoring area, it can directly impact the data collection and the outcome failures types. Thus, the main challenge is to extract significant information from raw data. In this way, outliers detection techniques have been gaining more and more prominence, contributing to increasing the data collection quality, and the improvement of these techniques (with the addition to outliers identification and treatment methods), may be used to detect defective sensors, filter noisy data, and discover relevant events. This thesis proposal is an architecture to detect, identify and treat outliers in a large scale WSN. One of the architecture modules is responsible to associate the outliers detection method, based on statistics, to a machine learning approach to identify and classify outliers, based on a time-space data correlation, and then, properly treat it. Another fundamental module of the proposed architecture is in charge to integrate techniques of data clustering with a

network cluster-tree topology dynamic formation strategy. To evaluate the proposed architecture, several simulated configurations were performed. The results show that the proposed architecture, with the associated modules, is able to adapt itself to the environmental changes in the monitored area, dynamically modifying its communication topology, which allows an accurate outliers detection, and consequently, a responsive detection to the relevant events that occur during monitoring.

Keywords: Event Detection, Outlier detection, Monitoring strategy, Cluster-tree based on data;;

Resumo

As Redes de Sensores Sem Fio (RSSF) surgem como uma importante área de pesquisa, incorporando trabalhos relacionados às áreas de comunicação, sensoriamento e tratamento de dados. Os cenários de uso desse tipo de rede são vastos, podendo ser citadas as aplicações de monitoramento ambiental, industrial, hospitalar, dentre outras. Nesse contexto, o cenário deste estudo é sobre o monitoramento de grandes áreas, as quais possuem especificidades desafiadoras, tais como grande área de cobertura espacial, variação na quantidade de sensores por área, dificuldade de implantação, e geração de grande quantidade de dados anômalos. Os recursos computacionais limitados dos nodos, associados à instalação em possíveis ambientes hostis e mudanças dinâmicas nas condições do ambiente monitorado impactam diretamente na coleta dos dados e nos tipos de falhas geradas. O principal desafio é o de extrair informações significativas, a partir dos dados brutos. Nesse sentido, técnicas para a detecção de *outliers* vêm ganhando cada vez mais destaque, contribuindo para melhorar a qualidade dos dados coletados. O aperfeiçoamento dessas técnicas, juntamente com técnicas para identificação e tratamento do *outliers*, pode ser usado para detectar sensores defeituosos, filtrar dados ruidosos e descobrir eventos relevantes. Nesta tese é proposta uma arquitetura para detecção, identificação e tratamento de *outliers* em RSSF de larga escala. Um dos módulos da arquitetura é responsável por associar métodos de detecção de *outliers* baseados em estatística a métodos de aprendizagem de máquina para identificar e classificar as anomalias, através da correlação

espaço-temporal dos dados e, em seguida, tratá-las adequadamente. Outro módulo fundamental da arquitetura é responsável por integrar técnicas de formação de *cluster* de dados com uma estratégia de formação dinâmica de topologia *cluster-tree* de redes. Para avaliar a arquitetura, diversas configurações de redes foram simuladas. Os resultados alcançados mostraram que a arquitetura, com seus módulos associados, foi capaz de se adaptar às mudanças de condições do ambiente monitorado, alterando dinamicamente a topologia da comunicação, permitindo uma detecção correta de *outliers* e, consequentemente, a detecção responsiva dos eventos relevantes que ocorram durante o monitoramento.

Palavras-chaves: Detecção de Eventos, Detecção de *Outlier*, Estratégia de monitoramento, *Cluster-tree* baseada em dados;

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

BI	Intervalo de Beacon
BO	macBeaconOrder
CAP	Período de Acesso com Contenção
CFP	Período sem Contenção
CSMA/CD	*Carrier Sense Multiple Access with Collision Avoidance*
CWA	Método Ponderamento por Confiança
FFD	*Full-Function Device*
IA	Inteligência Artificial
GPS	*Global Positioning System*
GTS	Intervalo de tempo garantido
IoT	Internet das Coisas
MTF	Média Tolerante a Falhas
RFD	*Reduced-Function Device*
RSSF	Redes de Sensores sem Fio
RSSI	Potência do Sinal Recebido
SD	Duração do Superframe

SO	*macSuperframeOrder*
SVM	Máquina de Vetor de Suporte

Sumário

1 Introdução e Contextualização

As Redes de Sensores sem Fio (RSSF) surgem como uma importante área de pesquisa, integrando trabalhos relacionados às áreas de comunicação e sensoriamento. Os desenvolvimentos obtidos a partir de seus estudos servem também como uma infraestrutura básica para outras áreas de pesquisa (YETGIN et al., 2017) (ALCARAZ et al., 2010). Os cenários de uso desse tipo de rede são vastos, variando desde as aplicações de monitoramento ambiental, hospitalar, agricultura inteligente, engenharias de estruturas, refinarias petrolíferas, dentre outras.

Essas redes são usualmente abordadas em contextos multi-disciplinares, incluindo sistemas de comunicação, sistemas embarcados, processamento de informação, sistemas distribuídos e processamento de sinais. Consequentemente existe, uma grande variedade de áreas de investigação, por exemplo, detecção de anomalias, protocolos de roteamento, estratégias de localização, estratégia de monitoramento, *design* de *hardware*, mineração de dados, processamento de informações e segurança.

Uma RSSF constitui-se, basicamente, de dezenas a milhares de pequenos nodos com baixa capacidade de processamento, reduzido espaço de endereçamento de memória e com sensores de baixo custo. Estes são utilizados,

individualmente, para monitorar e coletar dados do ambiente, que são enviadas através de comunicação sem fio para um ponto central (estação base). Usualmente, nos nodos também há restrições quanto ao consumo energético e o alcance da comunicação sem fio (MESMOUDI; FEHAM1; LABRAOUI, 2013; CHENG et al., 2012a; AKYILDIZ et al., 2002). A confiabilidade desse tipo de rede é obtida através da cooperação entre seus nodos, sendo que a comunicação com a estação base pode ser feita diretamente ou através de nodos intermediários, os quais passam também a cumprir o papel de roteadores na rede.

A característica da rede depende da aplicação, podendo esta ser de larga escala ou de pequenas dimensões; os nodos podem ser distribuídos em cada um dos casos de forma densa ou esparsa, o que é caracterizado pela quantidade de sensores implantadas por m^2 ou m^3. Os nodos distribuídos em grandes áreas geográficas precisam interagir uns com outros de forma autônoma. Essa autonomia, com relação à estação base e aos outros nodos passa a ser um importante desafio na implementação dessas redes, as quais são suscetíveis a ataques maliciosos e são vulneráveis a falhas (VELMANI; KAARTHICK, 2015; KIM; BANG; LEE, 2014; BHOJANNAWAR; BULLA; DANAWADE, 2013; JURDAK et al., 2011).

O emprego de RSSF no monitoramento de grandes áreas pode gerar uma grande quantidade de dados, e um dos desafios está relacionado com a extração de informações significativas a partir dos dados brutos. Nesta perspectiva, segundo Hall, Member e Llinas (1997), “a fusão da informação

cumpre o importante papel de fazer a combinação de dados sensoriais ou dados derivados dos dados sensoriais, tal que a informação resultante seja, em algum sentido, melhor do que seria possível caso essas fontes fossem usadas individualmente". Desta forma, a fusão da informação contribui para uma maior confiabilidade na análise, na robustez, na otimização e na transmissão de dados.

A fusão da informação necessita de dados íntegros para que seus resultados sejam confiáveis. Portanto, outro desafio relacionado ao monitoramento de grandes áreas por RSSF é o da identificação de *outliers*. Segundo Zhang, Meratnia e Havinga (2010) *outliers* ou anomalias são definidas como "observações que não correspondem a uma noção bem definida de comportamentos normais". As causas usuais das anomalias em RSSF são erros, mau funcionamento e ataques maliciosos aos nodos.

A detecção de *outliers* apresenta um papel fundamental para três importantes tarefas (i) identificar sensores defeituosos; (ii) filtrar dados ruidosos; e (iii) detectar eventos relevantes na área monitorada. Para cumprir essas tarefas, um aspecto importante a ser ressaltado na detecção de um *outlier* está na diferenciação deste entre evento relevante e dado espúrio. De uma maneira sucinta, quando um dado é detectado como *outlier* este passa por um processo de análise para determinar se é um dado espúrio (precisa ser retirado) ou se existe correlação espaço-temporal com dados oriundos de nodos vizinhos, definindo-o como evento relevante. As principais diferenças entre a detecção de dados espúrios e detecção de eventos relevantes encontradas na literatura são

discutidas, principalmente, pelos seguintes autores: (ZHANG et al., 2012; RASSAM; ZAINAL; MAAROF, 2013a; KIM; BANG; LEE, 2014). Resumidamente destacam-se:

- Técnicas de detecção de dados espúrios são independentes da semântica das aplicações, enquanto as técnicas de detecção de eventos, comumente, exploram conhecimentos prévios dos eventos que causam as anomalias;
- Técnicas de detecção de dados espúrios visam a identificação de anomalias através de estimação e comparação de medições, enquanto as técnicas de detecção de eventos comparam as leituras de sensores com padrões pré-definidos ou correlacionam os dados de nodos vizinhos de forma espaço-temporal;
- Dados espúrios resultantes de sensores defeituosos ou de ruídos no processo de comunicação são geralmente aleatórios e independentes por natureza, ao passo que os eventos podem ser correlacionados de forma espaço-temporal.

Nesse contexto, o cenário de aplicação desta pesquisa de doutorado está relacionado ao monitoramento de grandes áreas utilizando RSSF. Esse cenário possui especificidades desafiadoras tais como, área de cobertura e limite no alcance dos rádios dos nodos, variação na densidade (quantidade de sensores/área), distribuição irregular dos nodos, dificuldade de implantação, erros na transmissão dos dados, geração de dados anômalos, consumo energético da bateria e atrasos variados das mensagens devido à comunicação *multi-hop*.

Outro aspecto relacionado à significância dos dados capturados é o fato que, em áreas muito extensas, o valor de um dado específico pode não representar verdadeiramente o estado daquele ambiente monitorado. Grandezas medidas em áreas diferentes podem ter significados diferentes. Esse fato pode ocorrer, por exemplo, quando altas temperaturas são detectadas, sendo este um evento dado como normal para uma região seca, mas anômala para região úmida. Portanto, observa-se que os trabalhos voltados à detecção de eventos utilizam estratégias distintas quando comparados com trabalhos voltados à detecção de dados espúrios.

Finalmente, em redes de larga escala, um *outlier* pode ser associado a algum evento relevante que começou a ocorrer na área monitorada. As características do monitoramento, a taxa de transmissão dos dados e a própria topologia da rede deveria se adaptar a essa situação. Esses tipos de eventos geralmente são associados a alarmes que precisam ser disparados quando patamares específicos são alcançados pelos valores monitorados. Nesse ponto, foi observada uma lacuna nos trabalhos da literatura, os quais desconsideram o dado monitorado como um parâmetro para a construção dinâmica da topologia da rede.

O foco desta tese está em uma arquitetura para detecção, identificação e tratamento de *outliers* em RSSF de larga escala, com aplicação de métodos com baixo custo computacional, adequados a serem empregados a esse tipo de rede. A arquitetura incorpora uma estratégia de formação dinâmica de redes *cluster-tree* baseada em dados. A ideia básica é que a topologia se reconfigura dinamicamente,

buscando priorizar o monitoramento de áreas onde possam ocorrer eventos relevantes, com impacto na redução dos atrasos fim a fim dessas áreas.

1.1 Motivação

Aplicações de monitoramento geralmente fazem tomadas de decisão com base nos dados sensoriados do ambiente recebidos através de uma estação base. Contudo, em RSSF em larga escala, os dados estão sujeitos a vários fatores que propiciam a geração de *outliers*. Segundo Hodge e Austin (2004), as principais causas estão relacionadas com dispositivos imprecisos, esgotamento da bateria, ambientes hostis sujeitos a manipulações e redes de sensores com grande quantidade de sensores, o que pode gerar uma maior quantidade de erros. Nesse sentido, o processo de reconhecer anomalias é muito importante para análise dos dados, etapa necessária na identificação de valores discrepantes (RASSAM; ZAINAL; MAAROF, 2013a; AKYILDIZ et al., 2002).

A detecção de *outliers* em RSSF é o processo de identificar os dados que se afastam do padrão com base em um modelo ou estimação do mesmo. As observações cujas características diferem significativamente do perfil normal são considerados como *outliers* (ZHANG et al., 2012). Muitos fatores evidenciam a necessidade do aperfeiçoamento e evolução de técnicas de detecção de *outlier*. O processo de identificação do *outlier* resulta em duas possibilidades: (i) o dado é considerado como espúrio, e consequentemente, descartado, ou (ii) o dado é classificado como um evento relevante, precisando geralmente receber atenção do sistema de monitoramento.

Importante ressaltar que nas abordagens para a detecção de evento é necessário conhecer os vizinhos próximos (e seus dados monitorados) do nodo que detectou um *outlier* para fazer uma correlação espaço-temporal entre os dados. Na literatura várias taxonomias foram propostas para a localização dos nodos em RSSF, a destacar: (SINGH; SHARMA, 2015), (JIN et al., 2015), (YAO et al., 2015), (ZENG et al., 2013), (PANWAR; KUMAR, 2012) e (CHENG et al., 2012b). Estas categorizam os métodos de localização em: autolocalização (estimação desconhecida) e alvo/origem (uso de *Global Positioning System*).

As técnicas para detecção de eventos apresentam desafios relevantes como:

- Dificuldade de estabelecer correlação espacial devido ao desconhecimento exato da posição do nodo;
- Dificuldade de estabelecer correlação temporal devido à inexistência de sincronização de relógios entre os nodos;
- Áreas de grandes dimensões podem ter regiões de medições com diferentes características em suas grandezas monitoradas, obrigando o estabelecimento de vários conjuntos com diferentes intervalos de monitoramento;
- O grande número de nodos pode aumentar quantidade de dados anômalos.

Essas condições específicas para a detecção do evento vêm atraindo muitas pesquisas na busca de esquemas para

realizar a detecção de eventos utilizando nodos de baixo custo em RSSF de larga escala. Ou seja, a busca de esquemas que não necessitem de implantação de nodos com posição conhecida ou uso de *Global Positioning System* (GPS) ou outros dispositivos adicionais.

No contexto das RSSF, as propriedades desejáveis, de modo geral, para os métodos de detecção de anomalias devem considerar: modo de operação *online*, estrutura distribuída, adaptativa a mudanças, com mecanismos que busquem reduzir a dimensão dos dados e que explorem as correlações espaço-temporais dos dados (MERATNIA; HAVINGA, 2010; BHOJANNAWAR; BULLA; DANAWADE, 2013). O principal objetivo da detecção de anomalias em RSSF, portanto, é identificar valores discrepantes nos dados de forma distribuídas e *online* com alta precisão de detecção, mantendo o consumo de recursos computacionais e de redes mínimos (HODGE; AUSTIN, 2004).

Outro aspecto essencial sobre a detecção de eventos relevantes em RSSF de larga escala é a necessidade de se enfatizar o monitoramento nas regiões relacionadas ao evento, correlacionando esse problema a questões ligadas à comunicação da rede. Nesse cenário, podem-se ressaltar desafios quanto aos critérios utilizados para formação de topologias e rotas na rede, e na estratégia para a redução do atraso fim-a-fim na comunicação da rede para que o evento seja detectado o mais rapidamente possível na estação base. Esse desafio é exacerbado pelas condições dinâmicas usuais em RSSFs, pois falhas intermitentes nos nodos e mudanças constantes nos dados monitorados podem demandar o

estabelecimento de uma estratégia que preveja mudanças dinâmicas na topologia de comunicação para se ajustar. Em síntese, os principais problemas abordados nesta tese de doutorado podem ser resumidos pelas seguintes questões de pesquisa:

- Como realizar, em RSSF de larga escala, a detecção, identificação e tratamento de *outliers* utilizando nodos de baixo custo?

- Como desenvolver, em RSSF de larga escala, uma estratégia de formação de topologias *cluster-tree* baseadas em dados, com efeitos concretos no tempo de detecção de eventos relevantes na rede, mantendo um balanceamento na questão energética e no volume de tráfego gerado?

1.2 Objetivos

No sentido de responder as duas perguntas de pesquisa deste trabalho, o principal objetivo desta tese é desenvolver uma arquitetura que integre as principais etapas necessárias para detecção, identificação e tratamento de *outliers* em RSSF de larga escala. A arquitetura proposta é composta por dois grandes módulos, nomeadamente:

(a) Estratégia de agrupamento, seleção, e filtragem de dados; e

(b) Estratégia de formação de topologia em redes *cluster-tree* baseada em dados.

No módulo para "Estratégia de agrupamento, seleção e filtragem de dados" propõe-se associar métodos de detecção de *outliers* estatísticos a métodos de estimação do alcance dos nodos baseados em aprendizagem de máquina.

O módulo "Estratégia de formação de topologia em redes *cluster-tree* baseada em dados" adiciona à arquitetura um esquema de realimentação em malha-fechada, no qual mudanças significativas nos dados monitorados podem disparar um mecanismo de mudança de topologia, com o principal objetivo de detectar eventos relevantes na rede de forma eficiente e com menores atrasos fim-a-fim.

Objetivos Específicos

Dentre os elementos desta tese que são essenciais para atingir o objetivo geral, pretende-se alcançar os seguintes objetivos específicos:

- Estabelecer um método de estimação na localização dos nodos baseados na intensidade do sinal da comunicação dos nodos (RSSI), ou seja, sem a necessidade de dispositivos dotados de GPS ou de conhecimento prévio da localização dos nodos;
- Avaliar a integração entre os métodos de detecção de *outliers* baseado em estatística e o método de estimação da localização dos nodos em aprendizagem de máquina não supervisionado;
- Avaliar uma estratégia de formação de *cluster-tree* baseada em dados no processo de monitoramento, com

foco na redução dos atrasos fim-a-fim na detecção dos eventos relevantes na rede e no consumo de energético dos nodos.

1.3 Escopo do trabalho

Apesar de haver algumas propostas recentes relacionadas ao contexto das RSSF, técnicas estatísticas para detecção de *outliers* são estudadas há mais de um século. Não é propósito deste trabalho contribuir com uma nova técnica estatística para detecção e/ou identificação de *outliers*. Nesse sentido, neste trabalho, a contribuição mais próxima a essa questão, é a de integrar técnicas leves, aplicáveis em RSSF, através de uma arquitetura que ajusta a topologia da rede dinamicamente, conforme os próprios valores dos dados monitorados.

Com relação à topologia, apesar de haver estudos sobre alternativas para formação da rede, tais como topologia em malha (*mesh*), o foco deste trabalho está na topologia *cluster-tree* por ser a mais eficiente energeticamente. Topologias em malha possibilitariam a existência de mais de um caminho de comunicação entre os nodos, aumentando a tolerância a falhas da rede. Contudo, este trabalho não tem como objetivo analisar topologias alternativas à *cluster-tree*.

Em redes IEEE 802.15.4, fortemente associada à formação de topologia *cluster-tree*, existe a necessidade de se estabelecer um escalonamento da comunicação (escalonamento de *beacons*). A escolha adequada de valores durante o escalonamento de *beacon* permite que a rede seja mais eficiente, em termos de economia energética (permitindo aos nodos

dormirem em tempos em que não são necessários); e em termos de atrasos da rede, permitindo que os *clusters* no caminho entre o evento e a estação base estejam ativos e repassem rapidamente os dados monitorados. Contudo, neste trabalho, não existe o objetivo de se comparar diferentes técnicas de escalonamento de *beacon*.

Finalmente, apesar do reconhecimento da necessidade da existência de mecanismos de segurança na rede para evitar ataques, tais como o de negação de serviço (*deny of service*), não é propósito dessa tese desenvolver, utilizar ou avaliar qualquer mecanismo de segurança.

1.4 Contribuições científicas

Esta tese apresenta uma série de etapas relacionadas à proposta tema, que envolve: (i) o desenvolvimento de uma estratégia de agrupamento, seleção e filtragem de dados orientada a detecção de eventos; (ii) o desenvolvimento de uma estratégia de formação para redes *cluster-tree* baseada em dados; e (iii) sua implementação e validação em simulador de RSSF. Os esforços para planejar, desenvolver e realizar todas essas etapas no contexto de RSSF é um dos grandes avanços desta pesquisa.

A proposta permite a utilização de diferentes métodos em diferentes estágios da arquitetura, de acordo com o objetivo da RSSF. Por exemplo, aplicar outros métodos para agrupamento de dados; ou utilizar outros métodos para detecção de *outliers*, ou desenvolver outras estratégias para formação em redes *cluster-tree*. Especificamente, tendo em vista os objetivos e alcance, esta tese apresenta as seguintes

contribuições:

1. Concepção de uma arquitetura voltada à detecção, identificação e tratamento de *outliers*;

2. Desenvolvimento de uma estratégia inovadora para formação de redes *cluster-tree* baseada em dados;

3. Validação e apresentação de resultados dos métodos de agrupamento, seleção e filtragem associados à estratégia de formação de redes baseada em dados;

4. Implementação da estratégia de formação para redes *cluster-tree* baseada em dados no simulador Castalia para realização de pesquisas futuras;

1.5 Organização do Texto

Esta tese de doutorado está organizada da seguinte forma. O Capítulo 2 trata sobre os conceitos e definições fundamentais ao contexto da pesquisa, destacando-se os relacionados a RSSF e a fusão da informação. O Capítulo 3 aborda os princípios, os desafios e os requisitos para classificação dos métodos de detecção de *outliers*, e também apresenta uma revisão da literatura sobre os métodos de detecção de *outliers* associadas a detecção de eventos no contexto de RSSF. O capítulo finaliza com uma análise de trabalhos relacionados sobre métodos de detecção estatísticas e de estimação de alcance para RSSF. O Capítulo 4 descreve a arquitetura em termos de contribuições para detecção de eventos, problemas específicos da arquitetura e elementos para comparações desta com outras abordagens que têm a finalidade

na detecção de eventos. Ademais, detalha a arquitetura para detecção de eventos em RSSF de larga escala, sendo descrita cada etapa que a compõe. Para melhor detalhamento e discussão dos resultados, a arquitetura foi dividida em duas estratégias a seguir. O Capítulo 5 apresenta a avaliação e resultados obtidos na estratégia de agrupamento, seleção e filtragem dos dados. Na sequência, o Capítulo 6 apresenta a avaliação e resultados obtidos na estratégia de formação de topologia de redes baseada em dados. Por fim, o Capítulo 7 traz as considerações finais desta tese e discute as possíveis linha de trabalhos futuros.

2 Internet das Coisas, Redes de Sensores Sem Fio e Fusão da Informação

Neste capítulo apresentam-se conceitos e definições relacionados à Internet das Coisas, RSSF e fusão da informação. A Internet das Coisas seus conceitos, paradigmas e desafios são contextualizados na Seção 2.1. A Seção 2.2 apresenta uma visão geral sobre RSSF, seus conceitos fundamentais e aplicações. Nessa seção, são tratados, especificamente, os aspectos de comunicação em RSSF, escalonamento de sono dos seus nodos, e protocolos de roteamento RSSF. Na Seção 2.3 são discutidos aspectos relacionados aos conceitos, definições e classificação da fusão da informação. As considerações finais do capítulo são colocadas na Seção 2.4.

2.1 Internet das Coisas

A diversidade de cenários e a potencialidade das RSSFs propiciam a evolução para um novo conceito, chamado de Internet das Coisas (IoT), no qual cada objeto na vida humana será equipado com sensores que se comunicam uns com os outros, através de uma rede (ATZORI; IERA; MORABITO, 2010). Espera-se que a IoT constitua a Internet do futuro, formando uma rede mundial de objetos interligados,

endereçáveis com base em protocolos padrão de comunicação EPoSS (2008).

No paradigma da IoT, os sensores são a base da infraestrutura de sensoriamento e um dos elementos mais importantes é a RSSF, atuando como infraestrutura que fornece meios para acessar informações sobre o mundo físico por qualquer sistema computacional Alcaraz et al. (2010). Para possibilitar a integração das RSSFs com a IoT, diferentes tecnologias são desenvolvidas como:

- padrão 6LoWPAN [RFC 4944], definido pelo IETF e que permite a transmissão de pacotes IPv6 em redes computacionais;
- protocolo Bluetooth [RFC 802.15.1], usado por dispositivos de comunicação com restrição de energia;
- protocolo RPL (*Routing Protocol for Low-Power and Lossy Networks*) [RFC 6550], para a tarefa de roteamento;
- protocolo CoAP (*Constrained Application Protocol*) [RFC 7252], um substituto para o HTTP da camada de aplicação. É um protocolo orientado a serviços projetado para redes de objetos inteligentes.

Apesar da existência desses padrões que facilitam a integração com a IoT, o principal desafio em uma RSSF é extrair conhecimento de alto nível a partir de dados brutos. Recentemente, o tema de detecção de *outliers* em RSSF vem ganhando destaque por fornecer meios eficientes na procura de valores que não seguem o padrão normal dos dados, atribuindo

confiabilidade na identificação de dados anômalos, robustez na análise de dados e redução no tráfego de dados espúrios (ZHANG; MERATNIA; HAVINGA, 2010). Várias áreas pesquisam sobre detecção de anomalias, nomeadamente estatística, mineração de dados, teoria da informação e decomposição espectral. Detecção de anomalias também é tema em vários domínios de aplicações como invasão de redes, análise de desempenho, detecção de fraudes e previsão do tempo.

Apesar da vasta área de aplicação, a finalidade principal no uso de RSSF pode ser sintetizada como "fornecer informações relevantes para tomar decisões em tempo hábil". Neste sentido, a análise de dados dos sensores tem fundamental importância na reflexão do verdadeiro estado monitorado. Entretanto, a medição de dados brutos pode ser imprecisa e não confiável. Essa situação é exacerbada em redes de grande porte (HODGE; AUSTIN, 2004). A imprecisão pode ser provocada pelo próprio dispositivo ou pelo ambiente instalado. As limitações em termos de recursos de armazenamento, energia, largura de banda e processamento podem contribuir para a geração de anomalias nos dados. As questões relacionadas à implantação dos nodos podem resultar em alteração nos dados. Ainda neste sentido podem haver ataques maliciosos ou destruição dos sensores, podendo ocasionar baixa na qualidade e imprecisão dos dados (AKYILDIZ et al., 2002).

Existe uma gama de soluções para detecção de anomalias em redes tradicionais. Entretanto, essas soluções não podem ser portadas diretamente para RSSF pelas limitações em termo *hardware* e consumo energético. As técnicas para detecção de anomalias para redes tradicionais se concentram na

camada de rede, portanto o desenvolvimento de técnicas adequadas para RSSF para camadas mais baixas são necessárias (FAROOQI; KHAN, 2012).

As soluções para detecção de anomalias para RSSF devem ser caracterizadas pela eficácia e eficiência na utilização de recursos limitados. A eficácia na detecção de anomalias é representada pela exatidão, taxas de detecção e alarmes. A eficiência na detecção de anomalias é representada pelo consumo de energia e utilização de memória. Portanto, a proposta de soluções para detecção de anomalias deve considerar a melhor eficácia na detecção, além de consumir menor quantidade de recursos de armazenamento, processamento e consumo energético (ZHANG; MERATNIA; HAVINGA, 2010).

2.2 Redes de Sensores sem Fio

As RSSFs surgem como uma área crescente de pesquisas devido ao grande potencial de sua utilização em várias áreas. Entre as funções dos nodos que compõem uma RSSF estão a detecção de anomalias no ambiente monitorado, e o processamento e transmissão dos dados para uma estação base. Por outro lado, existem restrições quanto ao consumo energético já que a substituição de baterias pode ser um processo dispendioso (RASSAM; ZAINAL; MAAROF, 2013a).

Um nodo é composto, em sua estrutura básica, por quatro componentes principais: unidade de processamento de dados, unidade de alimentação, unidade de monitoramento e unidade de transmissão de dados (Figura 1).

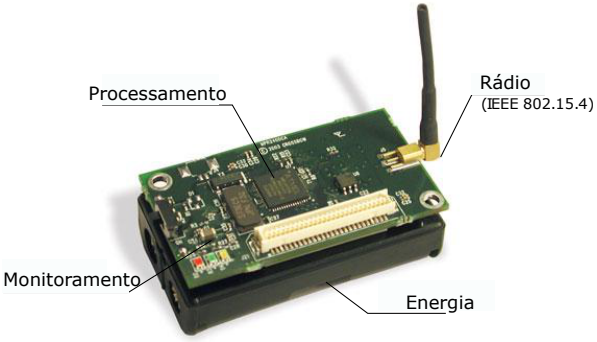


Figura 1 – Estrutura do Nodo, Adaptada de (YICK; MUKHERJEE; GHOSAL, 2008)

As RSSFs possuem desafios em relação a outras redes. O primeiro é relacionado a não determinação do posicionamento dos nodos no ambiente gerando, consequentemente, há muitos esforços na concepção de algoritmos e protocolos para auto-localização. O segundo é relacionado à distribuição no processamento de dados, promovendo o processamento local de dados simples nos nodos e o envio dos dados processados para a estação base (AKYILDIZ et al., 2002). O terceiro é com relação à energia, sendo, três as fontes de seu consumo: sensoriamento, processamento e transmissão dos dados. Hill et al. (2000) aponta que a maior parte da energia é gasta com a transmissão e recepção dos dados, na proporção de um bit transmitido para milhares de operações na unidade central de processamento do sensor.

Quanto ao tipo de sensores em RSSF, estes monitoram uma grande variedade de grandezas físicas, tais como: umidade,

luminosidade, pressão, ruído acústico, temperatura, velocidade, entre outras (ESTRIN et al., 1999). Os dados são usualmente monitorados de forma contínua podendo gerar grandes volumes de dados coletados. Quando uma RSSF monitora apenas um tipo de dado, como pressão ou temperatura, a este é atribuído o nome de dado univariado. Em RSSF que monitoram diferentes tipos de dados, simultaneamente, esses dados recebem o nome de dados multivariados (RASSAM; ZAINAL; MAAROF, 2013a).

Os autores em (YICK; MUKHERJEE; GHOSAL, 2008) categorizaram as RSSF em: monitoramento do ambiente e o rastreamento de objetos. As áreas de aplicação para dessas redes envolvem monitoramento em negócios, automação, ambientais, hospitalares, nas engenharias, saúde entre outros diversos cenários.

2.2.1 Topologia

Os dispositivos para redes IEEE 802.15.4 são classificados de acordo com a função que executam na redes (Zigbee Alliance, 2011):

- *Reduced-Function Devices (RFDs):* são dispositivos com função reduzida, mais limitados em termos de recursos de memória e processamento. Geralmente utilizados na borda da rede, não atuando no reenvio de mensagens e nem como coordenador. Sua comunicação é unicamente com dispositivos completos descritos no próximo item.

- *Full-Function Devices (FFDs):* são dispositivos com função completa e que podem atuar em toda topologia da rede podendo ter função de coordenador ou roteador. Por

serem dispositivos com mais recursos computacionais são considerados completos.

A formação de uma RSSF pode conter vários dispositivos FFD e RDF. A natureza da aplicação determina o funcionamento da topologia da rede em: ponto-a-ponto *(peer-to-peer)* ou estrela *(star)* (Figura 2).

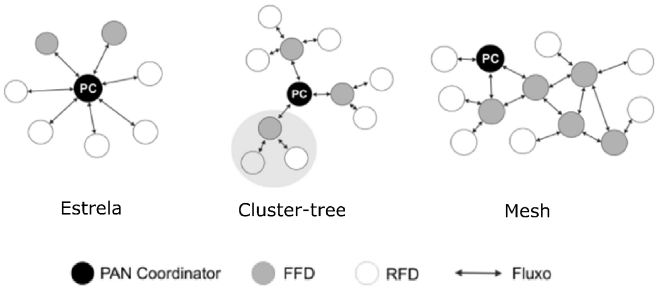


Figura 2 – Topologia em RSSF.

Na topologia estrela temos apenas um coordenador centralizado com mais recursos computacionais e alimentação de energia diferenciada, e os demais dispositivos são RFD. Todos dispositivos conectados ao coordenador utilizam um endereço para comunicação, permitindo que redes estrelas funcionem em conjunto e independentes.

Na topologia ponto-a-ponto todos os dispositivos podem executar função de roteador podendo, esta ter múltiplos saltos para enviar a mensagem entre os nodos até a estação base. Essa formação de topologia permite variações mais complexas, tais como redes em árvore *(cluster-tree)* ou em malha *(mesh)*. Uma das grandes vantagens da topologia *cluster-tree* é que qualquer

dispositivo FFD pode ser o coordenador, também chamado de *cluster-head*, provendo a sincronização da redes.

2.2.2 Aspecto da Comunicação nas RSSF

Os aspectos da comunicação em RSSF estão diretamente relacionadas ao tipo de topologia adotada. Duas das topologias mais usadas em RSSF são a *mesh* e *cluster-tree*.

Na topologia *mesh* a comunicação é descentralizada e permite salto através dos nodos vizinhos, os quais podem atuar como roteadores e encaminhar pacotes para outros nodos que não estão diretamente no alcance do nodo destino (AKYILDIZ; WANG; WANG, 2005). Dentre as vantagens da topologia *mesh* podem-se destacar (PANWAR; KUMAR, 2012) a baixa complexidade, flexibilidade, redundância no roteamento e boa escalabilidade. Entretanto, a topologia *mesh* é omissa no mecanismos de sincronização de tempo, o que possibilitaria configurar diferentes ciclos de trabalho nos nodos a fim de aumentar o tempo de vida da rede (PAN; TSENG, 2008). Este fato em especial é um complicador para uma RSSF, já que a eficiência energética é um dos seus aspectos fundamentais.

A topologia *cluster-tree* é indicada como mais adequada para redes de grandes escala (LI et al., 2011), nos quais os nodos são agrupados em *clusters* coordenados por um *cluster-head*. O *cluster-head* é responsável pela sincronização entre os nodos que fazem do seu agrupamento e também pela comunicação com os demais *clusters-heads*. Nesta topologia o coordenador principal, raiz da árvores, é responsável por gerenciar todas as sincronizações e atividades da redes.

A topologia *cluster-tree* é uma variação da topologia

ponto-a-ponto que utiliza como controle o modo *beacon* ativado. Nesse modo, as atividades de comunicação realizadas utilizam uma estrutura chamada de *superframe*. Estes *superframes* são transmitidos periodicamente por nodos *clusters-heads*. O controle da sincronização dos nodos associados se da através do *beacon*, a Figura 3 apresenta o *superframe* do IEEE 802.15.4 (Zigbee Alliance, 2011).

A estrutura de *superframe* é composta por duas partes: período ativo e inativo. No período inativo, os nós podem entrar no modo de baixa energia para economizar energia. No período ativo, a comunicação entre nodos do *cluster* pode ser realizada. A parte ativa compreende dois períodos: período de acesso com contenção (CAP) e período sem contenção (CFP).

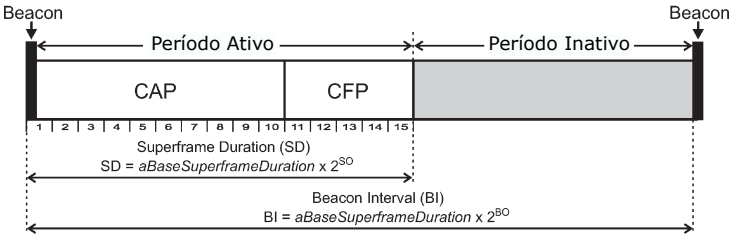


Figura 3 – IEEE 802.15.4 - Estrutura do superframe. Fonte: (Zigbee Alliance, 2011).

No Período de Acesso com Contenção (CAP), os nodos disputam o canal de comunicação usando o mecanismo de acesso ao meio *Carrier Sense Multiple Access with Collision Avoidance* (CSMA/CD). Para aplicações que requisitem baixa latência ou específica largura de banda, o período Período sem Contenção (CFP) é introduzido. No CFP, o nodo coordenador pode alocar intervalos de tempo garantidos (GTS) para

dispositivos específicos. Nesse slots, nodos podem transmitir quadros de dados sem disputar o acesso ao meio sem fio. O nodo coordenador pode alocar até sete Intervalo de tempo garantido (GTS) para seus nodos e cada um pode usar mais de um GTS.

A estrutura do superframe é definida por parâmetros macBeaconOrder (BO) e *macSuperframeOrder* (SO). Esses parâmetros definem o Intervalo de Beacon (BI) e a Duração do Superframe (SD), respectivamente. O Intervalo de Beacon (BI) define o intervalo que o coordenador deve transmitir seus quadros de *beacon*. A Duração do Superframe (SD) define o comprimento da parte ativa do superframe.

Embora o padrão IEEE 802.15.4 considere a topologia *cluster-tree*, este não fornece mecanismos ideais para trabalhar com redes *cluster-tree*. Por isso, a especificação ZigBee define a camada de rede e de aplicação na pilha de protocolos IEEE 802.15.4, fornecendo mecanismos que permitem a construção de redes *cluster-tree*, tais como de endereçamento e roteamento de hierárquicos. No entanto, a redes *cluster-tree* baseadas em IEEE 802.15.4/ZigBee ainda impõem várias questões desafiadoras, incentivando os pesquisadores a desenvolver novos protocolos e algoritmos para resolver questões de pesquisa abertas sobre a natureza típica de RSSF de larga escala (LEÃO et al., 2017).

2.2.3 Escalonamento de Sono

As atividades cíclicas de chaveamento entre períodos de energia ativo e inativo é conhecido com ciclo de trabalho ou *duty cycle*. Já o mecanismo que realiza o revezamento entre os

nodos nas tarefas de monitoramento e comunicação visando economia de energia é conhecido com escalonamento de sono ou *sleep scheduling* (Zigbee Alliance, 2011).

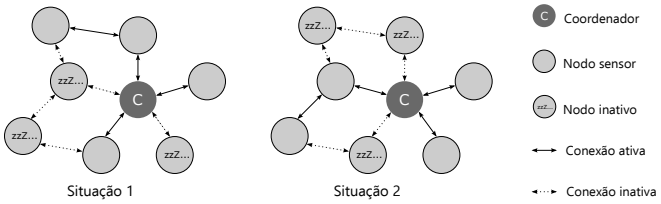


Figura 4 – Escalonamento de sono e transições de estados.

As questões sobre o projeto de escalonamento de sono geralmente envolvem outras questões, tais como estratégia de instalação dos nodos, capacidades dos nodos sensores, área de monitoramento, alcance de transmissão, sincronização de tempo, informações de localização e distância. Assim, deve-se observar o mecanismo de escalonamento apropriado para aplicação em questão.

As abordagens utilizadas para realizar o escalonamento de sono em RSSF podem ser classificadas em três tipos (NAKAMURA; LOUREIRO; FRERY, 2007), (LEAO et al., 2016): arbitrária, probabilista e síncrona.

Na abordagem arbitrária, o nodo coordenador decide quando o nodo sensor entra em modo inativo (modo *sleep*). Nesse caso, é necessário que o coordenador tenha conhecimento de todos os nodos conectados a ele. Assim é possível executar rotinas de seleção para realizar o escalonamento de sono. Como vantagem, pode considerar o estado atual das baterias dos nodos sensores. A desvantagem recai justamente na

centralização das decisões no coordenador.

As abordagem probabilistas preveem o escalonamento de sono através de métodos não deterministas, ou seja, os nodos entram em estado de baixo consumo energético em momentos aleatórios. O coordenador varre todo o conjunto de nodos ativos e de acordo com a probabilidade, decide sobre o estado futuro dos nodos. Uma desvantagem nesse tipo de abordagem é justamente desconsiderar a cobertura homogênea da rede. Entretanto, em algum tipo de aplicação isso pode ser suficientemente útil e vantajoso.

A abordagem síncrona utiliza métodos de sincronização para chaveamento entre os modos ativo e inativo dos nodos. Isto é, os nodos entram em acordo para decidir quais permanecem ativos e quais entram em modo *sleep* de forma que a cobertura da rede seja mantido em uma nível mínimo preestabelecido. Como vantagem, cite-se a consideração da cobertura espacial mínima da rede. Por outro lado, aumenta-se o número de mensagens trocadas entre os nodos da rede.

2.2.4 Protocolo de Roteamento para RSSF de Larga Escala

Esta seção aborda um aspecto de fundamental importância para RSSF de larga escala que consiste no processo de encontrar caminhos para a mensagem entre a origem e o destino, o roteamento. Para tal, o processo de identificação dos pares envolvidos e dos intermediários necessários é essencial para comunicação. Cabe ressaltar que a redução dos custos de fabricação dos nodos facilitou a implantação em múltiplas regiões de interesse e em grande quantidade, viabilizando RSSFs de larga escala (LI et al.,

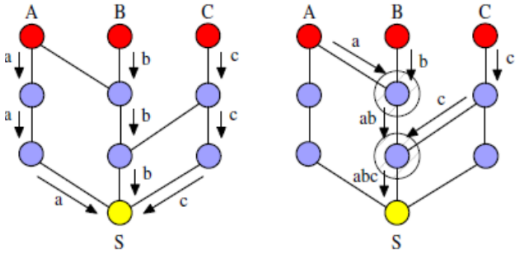
2011).

Em RSSF, o roteamento das mensagens tem diferentes formas de ser realizado, sendo que a eficiência da rede é influenciada pela forma como ocorre o roteamento. Uma das formas mais aplicada para RSSF em larga escala é a comunicação *multi-hop*. Esta forma permite utilizar os nodos intermediários como roteadores para alcançar o destino da mensagem. Neste sentido, é necessário entender as limitações das RSSF quanto ao consumo energético e a dificuldade de reposição da bateria. Por outro lado, formas de aperfeiçoar a comunicação, reduzindo o volume de dados transmitidos e economizando energia nas transmissões são desejáveis.

Neste contexto, a fusão de informação oferece uma opção para melhorar a qualidade dos dados, reduzindo as transmissões e consequentemente o consumo de energia. O conceito é realizar o processamento dos dados de forma local e/ou distribuído na rede, a fim de reduzir transmissões desnecessárias de dados redundantes, economizando energia (BAHREPOUR et al., 2009).

Deste modo, no processo de roteamento pode-se fundir e sintetizar os dados redundantes. Segundo Loureiro (2006), os tipos de roteamento em RSSF, Figura 5, podem ser: centrado em endereços ou centrado em dados.

Na Figura 5a os nodos A, B e C enviam dados para o coordenador S. No roteamento centrado em endereço, a transmissão desses dados acontece em 9 mensagens. No roteamento centrado em dados, Figura 5b, as mensagem são reduzidas para 6, nos quais alguns nodos destacados fazem a fusão da informação. Portanto, a utilização das abordagens



(a) Roteamento centrado em endereços.

(b) Roteamento centrado em dados.

Figura 5 – Tipos de roteamento em RSSF. Fonte: (LOUREIRO et al., 2003).

centradas em endereço ou em dados influencia no desempenho da rede (LOUREIRO et al., 2003).

Outro aspecto relacionado às RSSFs são que suas aplicações geralmente buscam extrair dados de um grupo de nodos ou região e não apenas de um nodo individual. Consequentemente, soluções de roteamento centrado em dados atendem melhor as restrições de RSSF.

Em RSSF de larga escala o número de nodos pode alcançar milhares e o objetivo da escalabilidade no algoritmo de roteamento é suportar longas distâncias, nas quais os dados monitorados podem viajar por grupos de nodos até alcançar o coordenador. De acordo com a estrutura da rede, os algoritmos de roteamento para RSSF recaem em três classes conhecidas como (LI et al., 2011):

- **Algoritmo de rede plana**: todos os nodos operam sobre as mesmas regras enviando os dados para o

coordenador, sendo possível o envio entre coordenadores de níveis diferentes até o coordenador principal;

- **Algoritmo de rede hierárquica**: todos os nodos são divididos em vários grupos com diferentes níveis de responsabilidade. Os nodos de nível alto são responsáveis por agregação e alguns por gerenciamento do trabalho. Os nodos de nível baixo são responsáveis pelo sensoriamento e coleta da informação;
- **Algoritmos baseado em localização**: utiliza informação da posição para enviar os dados para áreas de interesse. A localização é adquirida através de hardware específico adicionados aos nodos ou do conhecimento da posição dos nodos.

Nesta tese, foi utilizada a classe de algoritmo de rede hierárquica. Em comparação ao roteamento plano, a hierárquica propõe consumo energético mais eficiente, devido aos nodos serem tratados com regras diferentes, especialmente em termos de agregação dos dados e do controle de fluxo dos pacotes (AL-KARAKI; KAMAL, 2004). O algoritmo baseado em localização tem alto custo para RSSF de larga escala, por necessitar de adição de *hardware* específico.

Por ser tratar do algoritmo mais adequado para RSSF de larga escala, detalharemos os algoritmos de roteamento hierárquico a seguir.

2.2.4.1 Roteamento Hierárquico

No roteamento hierárquico são definidas duas classes para os nodos: coordenadores e os de sensoriamento. Os nodos

de sensoriamento coletam dados e transmitem para seu coordenador. Os nodos coordenadores podem transmitir dados para outros coordenadores de nível superior e também podem realizar sensoriamento com fusão da informação nas transmissões.

Os principais algoritmos desta classe são apresentados na sequência:

Protocolo LEACH: O protocolo LEACH *(Low-Energy Adaptive Clustering Hierarchy)* (HEINZELMAN; CHANDRAKASAN; BALAKRISHNAN, 2002) foi desenvolvido para redes homogêneas com o foco na redução do consumo energético. Esse protocolo opera formando agrupamentos e na escolha de um nodo líder *(cluster-head)*. O *cluster-head* tem a função de encaminhar os dados para a estação base em um único salto, limitando o tamanho da rede pelo alcance do rádio de transmissão. A rede é sincronizada inicializando todos os nodos ao mesmo tempo, sem especificar o grau da sincronização, podendo ocasionar início de novos ciclos em momentos desnecessários.

Protocolo LEACH-C: O esse protocolo é uma modificação do LEACH (LINDSEY; RAGHAVENDRA, 2002) por centralizar na estação base as decisões sobre a formação dos grupos. Com esta centralização ocorre uma distribuição mais eficiente dos grupos na rede. Quando a rede é inicializada, todos os nodos enviam informações da sua posição e do nível de energia para a estação base. Essas informações determinam a formação dos grupos e *(cluster-head)*, que funcionam na sequência como o LEACH tradicional.

Protocolo TEEN: O protocolo TEEN *(Threshold*

sensitive Energy Efficient sensor Network) (MANJESHWAR; AGRAWAL, 2001), assim como o protocolo LEACH são algoritmos de roteamento hierárquico; suas diferenças estão nos dados a serem transmitidos a cada intervalo de tempo. Os autores propõem classificar a rede em pró-ativa e reativa. A rede pró-ativa possui uma taxa constante de envio dos dados, enquanto que a reativa transmite dados quando há variação acima de um limite pré-definido. A estratégia da escolha do *cluster-head* é a mesma utilizada pelo protocolo LEACH, diferenciando-se na fase de transmissão dos dados. Nessa fase, adotam-se dois parâmetros para indicar a necessidade da transmissão dos dados: o *hard threshold* (imediato) e o *soft threshold* (no intervalo de alocação do nodo).

Protocolo PEGASIS: O protocolo PEGASIS *(Power eficience Gathering in Sensor Information Systems)* (LINDSEY; RAGHAVENDRA; SIVALINGAM, 2002) opera utilizando os nodos vizinhos mais próximos como roteadores até a estação base. O intervalo de tempo para os nodos vizinhos escolhidos é finito e rotativo. A ideia da estratégia é rotacionar os nodos utilizados como roteadores para equalizar o consumo da bateria e reduzir o volume de transmissões na rede. O protocolo assume algumas premissas para sua melhor aplicação como: nodos homogêneos, alcance direto dos todos os nodos a estação base e nível de energia uniforme.

Protocolo ICA: O protocolo ICA *(Inter Cluster Rounting Algorithm)* (ZHOU; CAO; GERLA, 2009) utiliza o protocolo LEACH como base, buscando melhor eficiência no consumo energético. O protocolo assume a posição geográfica da estação base e dos nodos da rede, após o *broadcasting*

enviado pela estação base. O ICA segue a mesma regra de formação de *(clusters)* do LEACH, diferenciando-se apenas na escolha pela proximidade do nodo ao *cluster-head* para associação.

2.3 Fusão de Dados

A terminologia relacionada à fusão da informação não possui uma nomenclatura unificada. Alguns sinônimos são conhecidos na literatura como, "fusão de sensores" por (ELMENREICH, 2002) "fusão de dados" (HALL; MEMBER; LLINAS, 1997) e integração de múltiplos sensores (LUO et al., 2002).

A fusão de dados é um sinônimo comumente empregado para fusão da informação. Contudo, para Elmenreich (2002), a fusão da informação tem visão mais ampla e abarca outros conceitos relacionados ao campo da fusão. Além disso, o termo fusão de dados pode ser aplicado também no contexto de fusão de dados brutos. Segundo Hall, Member e Llinas (1997), o termo fusão de sensores é apresentado como "a combinação de dados sensoriais ou dados derivados dos dados sensoriais, tal que a informação resultante é, em algum sentido, melhor do que seria possível quando essas fontes fossem usadas individualmente".

O termo fusão da informação para a *International Society of Information Fusion* (ISIF) é definido como "o estudo de métodos eficientes para automaticamente ou semiautomaticamente transformar informações de diferentes fontes e diferentes pontos no tempo em uma representação que forneça apoio efetivo para tomada de decisão humana ou

automatizada".

Como na literatura são encontrados vários termos relacionados à fusão, e os mais relevantes são fusão de dados e fusão da informação, para este trabalho adotaremos os dois de forma intercambiável. Em Dasarathy (1997) é abordada a comparação entre as técnicas relacionadas à fusão da informação, conforme apresentado na Figura 6. Fusão de sensor/multisensor é entendida como um subgrupo que opera com fontes sensoriais. A agregação de dados é visto como subgrupo que visa reduzir o volume de dados, que pode manipular qualquer tipo de dados/informação, incluindo sensores. A integração multisensor aplica a fusão da informação para fazer inferências usando dispositivos sensoriais e informações associadas para interagir com o ambiente. Desta forma, a fusão multisensor está na interseção de fusão da informação com a integração multisensor, conforme Figura 6.

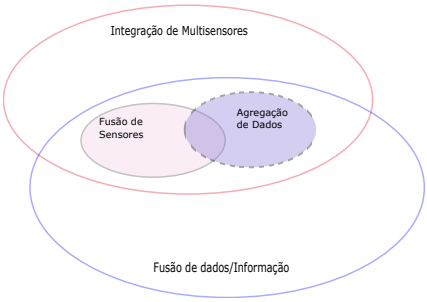


Figura 6 – Termos de fusão e seus relacionamentos. Adaptado: (DASARATHY, 1997).

As RSSFs possuem características peculiares para

implantação e medição dos dados. O ambiente de implantação pode estar sujeito a condições desfavoráveis como pressão, temperatura, interferência eletromagnética. Tais condições podem interferir na qualidade da medição. Por outro lado, mesmo em condições de implantação tidas como ideais, os sensores podem não apresentar medidas precisas. Este fato pode ocorrer por problemas de *hardware*, muitas vezes associados aos métodos utilizados para o monitoramento.

Outro desafio para RSSF está ligado à cobertura espaço-temporal do sensor. A cobertura temporal é relacionada com a capacidade de cumprir o propósito de monitoramento durante a vida útil, assegurando que não seja perdido nenhum evento relevante na rede. Os fatores de influência na cobertura temporal são: periodicidade da transmissão, volume de dados transmitidos e atraso de comunicação. A cobertura espacial está relacionada à área de alcance do sensor (NAKAMURA; LOUREIRO; FRERY, 2007).

Na tentativa de superar os problemas de coberturas espaço-temporais, limitações de *hardware* e falhas do sensor, o trabalho de Durrant-Whyte (1988) apontou na direção do uso de três propriedades: cooperação, redundância e complementariedade (detalhes serão vistos na Seção 2.3.1).

Neste contexto, a fusão de dados ganha espaço por aproveitar a capacidade de processamento do nodo para pré-processar dados de forma distribuída. No monitoramento, a utilização da abordagem centralizada com um único nodo é bastante comum. Entretanto, Elmenreich (2002) aponta desafios na utilização de uma único nodo:

- **Cobertura Temporal:** nodos precisam de um tempo "operacional" para realizar a medição e, posteriormente, para transmissão. Importante verificar se o tempo operacional é suficiente para observar a conclusão do evento;

- **Cobertura Espacial:** o nodo possui um limite de cobertura da área monitorada;

- **Disponibilidade:** perda da percepção do objeto monitorado, devido à perda de comunicação ou sensor defeituoso;

- **Incerteza:** um sensor é incapaz de reduzir a incerteza da medição devido a uma observação única sobre o evento;

- **Imprecisão:** medições de sensores individuais são limitados à precisão do próprio sensor utilizado.

Uma resposta apontada para as limitações existentes na utilização de um único sensor é a fusão de sensores. A ideia da fusão de sensores é aumentar a quantidade de sensores para conferir robustez às medições, independentemente da heterogeneidade dos sensores. As vantagens na utilização da fusão de sensores, apresentado em Elmenreich (2002) Hill et al. (2000) são:

- **Extensão da cobertura espaço-temporal:** vários nodos podem ampliar a área de cobertura e o tempo monitoramento do evento;

- **Robustez a falhas:** múltiplos sensores têm redundância inerente que permite o sistema fornecer informações, mesmo em caso de falha parcial;

- **Aumento de confiabilidade:** a medição de um sensor é confirmada por medições de outros sensores que cobrem o mesmo domínio;

- **Redução de ambiguidades e incertezas:** informações conjuntas reduzem o conjunto de interpretações ambíguas do mesmo domínio;

- **Robustez contra interferências:** aumento na dimensionalidade das medições do espaço do sistema;

- **Melhoria da resolução:** quando múltiplas medições independentes da mesma propriedade são fundidas, a resolução do valor resultante é melhor do que a medição de um único sensor.

O emprego de múltiplos sensores também confere maior área de cobertura espaço-temporal, mesmo que haja falhas parciais. Em termos de custo de infraestrutura, o nível de precisão da grandeza monitorada influencia no valor do sensor. Por outro lado, o uso de sensores de menor custo pode influenciar na perda da precisão. Em casos de captura um número de dados com má qualidade superior aos com boa qualidade, o desempenho global do sistema poderá ser comprometido (NAKAMURA; LOUREIRO; FRERY, 2007).

2.3.1 Classificação da Fusão da Informação

Esta seção aborda a classificação dos tipos de fusão conforme os níveis: entrada/saída, relação entre fontes e abstração.

Classificação Baseada na Entrada e Saída

Esse método de fusão foi proposto por Dasarathy (1997), que o classificou em cinco categorias de acordo com as entrada e saídas.

- **Dados Entram – Dados Saem (*Data In – Data Out*)**: nesta classe, a fusão da informação obtida como dados brutos e o resultado também como dados brutos, possivelmente mais precisos e confiáveis;

- **Dados entram – Características Saem (*Data In – Feature Out*)**: fusão da informação usa dados brutos para extrair características ou atributos que descrevem uma entidade. Aqui, a entidade significa qualquer objeto, situação ou abstração do mundo;

- **Características Entram – Características Saem (*Feature In – Feature Out*)**: fusão aplicada em um grupo de características para prover ou redefinir uma característica ou extrair novas;

- **Características Entram – Decisões Saem (*Feature In – Decision Out*)**: nesta classe, fusão pega um grupo de características da entidade e gera uma representação simbólica ou uma decisão;

- **Decisões Entram – Decisões Saem (*Decision In – Decision Out*)**: decisões podem ser fundidas a fim de obter novas decisões ou dar ênfase às anteriores.

As cinco categorias foram abstraídas em três níveis conforme Figura 7. O fluxo de informações evidencia a hierarquia das operações, sendo impossível utilizar dados brutos para tomar decisões.

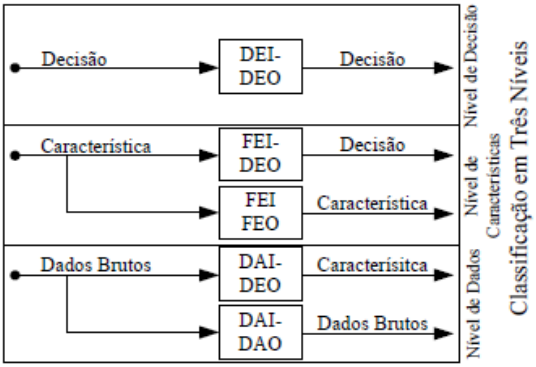


Figura 7 – Modelo Baseada em Entradas e Saídas. Fonte: (DASARATHY, 1997) e adaptado por: (CALLEGARO, 2014).

Classificação Baseada na Relação entre Sensores

De acordo com Durrant-Whyte (1988), baseada na relação entre os sensores, a fusão da informação classifica-se em:

- **Complementares:** os sensores não dependem diretamente um dos outros; existe uma relação de complementariedade do fenômeno observado. Na Figura 8 observa-se que os sensores S1 e S2 fornecem informações do objeto A e B. A complementariedade é o resultado da observação A + B;

- **Redundante:** se duas ou mais origens independentes fornecem a mesma parte da informação, estas partes podem ser fundidas para aumentar a confiança associada. A Figura 8 mostra que S2 e S3 fornecem informações do objeto B. Como resultados esta abordagem proporciona robustez, tolerância a falhas, exatidão e precisão, podendo ser utilizada com sensores heterogêneos e de forma competitiva. Esta fusão de dados para RSSF oferece maior qualidade evitando a transmissão redundante de informação;

- **Cooperativa:** uma rede cooperativa de sensores utiliza mais de um sensor para obter informações que não conseguiria se estivesse com apenas um. Neste tipo de fusão há perda de confiabilidade e precisão. Na Figura 8, S4 e S5 observam o mesmo objeto, mas as medições obtidas sobre o objeto C, não poderia ter sido obtida a partir de medições individuais.

É possível agrupar diferentes configurações de fusão em uma arquitetura híbrida. Como exemplo. em RSSF podemos monitorar a temperatura em uma mesma área geográfica com informações redundantes para conferir maior exatidão, fusão competitiva. Por outro lado, a fusão

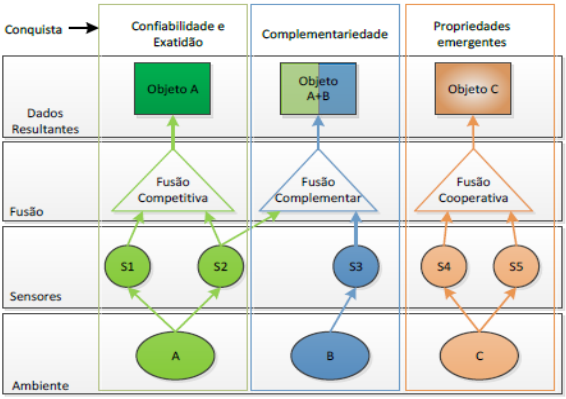


Figura 8 – Fusão competitiva, complementar, cooperativa. Fonte: (DURRANT-WHYTE, 1988) e adaptado por: (CALLEGARO, 2014).

complementar pode observar informações sobre umidade, temperatura e velocidade do vento gerando informações para previsão do clima. A fusão cooperativa pode ser empregada para unir dados de uma única grandeza.

Classificação baseado no Nível de Abstração

Em Luo et al. (2002), a fusão da informação é divida em quatro níveis de abstração: sinal, pixel, característica e símbolo.

- **Fusão Sinal (*signal-level*)**: fusão no nível de sinais com operação em sensores univariado ou multivariado. Pode ser utilizada em aplicações tempo real ou fase pré fusão. Exemplo de dados: temperatura, pressão e umidade;

- **Fusão de Pixel (*pixel-level*)**: fusão no nível de pixel trabalha com imagens onde cada pixel é utilizando para melhorar o processamento de imagem;

- **Fusão de característica (*feature-level*)**: a fusão no nível de característica extrai atributos de sinais ou imagens partir dos dados sensoriados;

- **Fusão de Símbolo *(symbol-level)***: a fusão de símbolo, utiliza informações antecipadamente conhecidas dos sensores e dos demais níveis para tomada de decisão.

Outra classificação para o nível de abstração foi proposta por (DASARATHY, 1997), composta por três níveis:

- **Fusão de baixo nível:** os dados brutos são fornecidos como entradas, combinadas em novos dados que são mais precisos do que as entradas individuais.

- **Fusão de médio nível:** é a fusão de dados brutos que representam características do objeto com melhor precisão. Geralmente utiliza reconhecimento extraídos de padrões usados em nível de decisão.

- **Fusão de alto nível:** fusão de nível de tomada de decisão, assume representação simbólicas como entrada e incorpora conhecimento antecipadamente para tomada de decisão.

2.4 Considerações do Capítulo

Este capítulo apresentou conceitos sobre IoT, a constituição básica do nodo e a grande variedade de cenários de

aplicações. Também se contextualizou sobre os protocolos de roteamento para RSSF necessários em rede de larga escala para comunicação dos dados. A questão da comunicação é importante devido ao consumo energético da bateria ativada nas transmissões. Outro aspecto é a utilização extensiva de sensores e o aumento na qualidade da precisão dos dados. Estes entendimentos possibilitam situar esta tese neste vasto contexto e relacioná-la a conceito de "Internet das Coisas"e suas expectativas futuras.

Este capítulo também foi envolveu conceitos relacionados à fusão da informação, igualmente sendo conhecida como fusão de dados ou fusão de sensores. Foram discutidas classificações para fusão da informação e suas características. O contexto da fusão da informação em redes de sensores é fundamentalmente importante, uma vez que RSSF utilizam a fusão para reduzir a quantidade e melhorar a exatidão da informação. Por fim, estes conceitos deram a base de sustentação para o desenvolvimento desta tese.

3 Outliers, Eventos e Clusterização de Dados

Este levantamento bibliográfico tem o foco na apresentação das técnicas para detecção, identificação e tratamento de *outliers*. No entanto, ao pesquisar sobre a detecção de *outliers* foi necessário entender os domínios conceituais e sua relação com a detecção eventos, detalhes na Seção 3.1. Na seção 3.2.1, contextualizamos sobre aprendizagem de máquina e sua relação essencial com arquitetura proposta. Os critérios para o estabelecimento da conectividade dos nodos e sua relação com nossa abordagem são apresentados na Seção 3.5. O processo de avaliação para distinção entre eventos e dados espúrios na Seção 3.2. As técnicas leves baseadas na abordagem estatística para detecção de *outliers* são descritas na Seção 3.3.3. Na Seção 3.6 apresenta-se uma análise dos principais trabalhos relacionados comparando-os com a proposta realizada nesta tese. Por fim, as considerações do capítulo são feitas na Seção 3.7.

3.1 Conceitos de Outliers

Conceitualmente, a detecção de eventos e de dados espúrios partem da mesma premissa: “medições que são significativamente diferentes do resto das observações ou padrão normal dos dados” (HODGE; AUSTIN, 2004). Portanto, após o processo de detecção de *outliers* é necessário

um processo de identificação, conforme Figura 9.

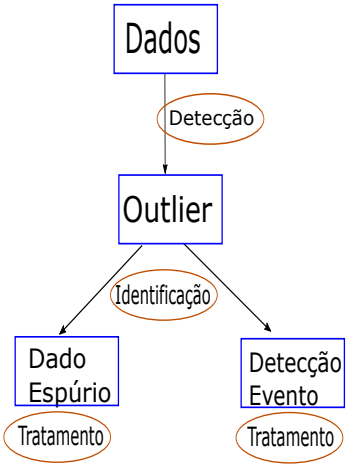


Figura 9 – Processos de detecção, identificação e tratamento de *outliers* em RSSF.

O termo *outlier* originalmente vem da área estatística (HODGE; AUSTIN, 2004). Uma definição clássica para *outlier* é de Hawkins no livro: "*Identification of Outliers*" de 1980, que definiu o termo como: "uma observação, que se desvia tanto de outras observações que desperta suspeitas de que foi gerado por um mecanismo diferente". Para Chandola, Banerjee e Kumar (2009), *outliers* são definidos como: "padrões nos dados que não se comportam com uma noção bem definida do comportamento normal". A detecção de *outlier*, definido em Hodge e Austin (2004) refere-se ao problema de encontrar padrões em dados que não correspondem ao comportamento estabelecido e esperado.

Neste contexto, vários fatores propiciam as gerações de *outliers* em RSSF (HODGE; AUSTIN, 2004), tais como: imprecisão dos dispositivos utilizados para o monitoramento; esgotamento da bateria dos nodos, implantação sujeita a manipulações danosas e número de sensores elevados possibilitando maior taxa de erros.

A identificação do tipo de fonte de anomalia é fundamental para tratamento adequado do *outlier* detectado. De forma ampla, os *outliers* podem ser agrupados em três categorias (JURDAK et al., 2011) (BAHREPOUR et al., 2009): de nodo, de rede e do dado.

Outlier do Nodo

São falhas exclusivamente no nodo, associadas a problemas de *hardware*, *software* e principalmente à questão da falha ou esgotamento da bateria. Nesse sentido pode haver uma subdivisão em: bateria, falha no nodo e reinicialização do nodo.

- **Bateria:** carga insuficiente na bateria ou falha no *hardware* da bateria. Carga insuficiente da bateria pode ser no nodo local ou do coordenador, por esgotamento da bateria ou em baterias recarregáveis potencialmente combinados aos fatores climáticos. A falha na bateria é associada ao não envio dos dados no tempo determinado, por esgotamento energético da bateria ou no caso de baterias recarregáveis ao limite de ciclos sendo necessária a substituição.

- **Falha no Nodo:** por componentes de *hardware* como memória, processador, transmissor, falha na integração

hardware/software e problemas na instalação danificando o nodo ou ambientes hostis.

- **Reinicialização do Nodo:** o operador pode definir um temporizador para reiniciar periodicamente os nodos por razões operacionais. Inesperadamente um erro de *software* altera o comportamento programado, fazendo que o comportamento pré-definido de reinicialização seja entendido como uma anomalia.

Outlier de Redes

Este tipo de *outlier* está relacionado ao problema de comunicação da RSSF como a inesperada variação da quantidade de pacotes na rede. As principais causas são:

- **Perda da conectividade**: quando há uma interrupção dos pacotes de dois ou mais nodos para um nodo específico. A perda de conectividade pode ocorrer individualmente, em grupos de nodos ou em toda a rede. A falta de recepção de pacotes em um determinado tempo rede pode indicar parcial ou total perda de conectividade com o nodo.

- **Intermitência na conectividade**: é a frequência de recepção de dados de um determinado nodo em relação ao coordenador com limite estável do enlace. Essa anomalia pode atingir individualmente ou grupos de nodos, sendo que esta detecção envolve criar um limite para taxa de entrega de pacotes ou variabilidade de qualidade do enlace.

- ***Loops* no roteamento**: ocorre quando um pacote é retransmitido em vários nodos e chega de volta ao nodo de origem. Essa anomalia é de difícil detecção, por isso seu processo de detecção envolve identificação de pacotes pela origem e conhecimento sobre a topologia da rede.

- **Tempestade de *broadcasting***: essa anomalia ocorre quando o nodo perde a conectividade de encaminhamento e envia pacotes de forma contínua para seus vizinhos para descobrir caminhos alternativos a estação base. Esta anomalia afeta todos os nodos da rede.

Outlier de Dados

Outliers são geralmente causados pelo *hardware* de sensores defeituosos, descalibrados ou variações no ambiente. Entretanto, é necessário entender que anomalias geradas por problemas nos sensores são consideradas como *outlier* de dados e não *outlier* de nodo, porque se manifesta em valores extremos ou inaceitáveis.

Por outro lado, alterações nos ambiente fazem com que os valores dos dados nos sensores mudem rapidamente, podendo ou não estar num limite razoável. Dessa forma, a distinção de anomalias de dados por comparação espaço-temporal a partir dos sensores é fundamental.

Existem três categorias para *outlier* de dados: temporal, espacial e a espaço-temporal (ZHANG; MERATNIA; HAVINGA, 2010):

- **Temporal:** exibem várias características, tais como: alta

oscilação nas leituras, que pode significar grandes mudanças em eventos detectados no ambiente; distorções das leituras graduais, que podem indicar a necessidade de recalibração do sensor. Leituras contínuas de mesmo valor por longo período de tempo, que podem indicar travamento do sensor. Leituras fora dos limites aceitáveis representam valores de sensores que fisicamente não são possíveis evidenciando um mau funcionamento do nodo. A detecção de *outlier* de dados temporais pode requerer um armazenamento de histórico individual localmente ou na estação base com os valores dos dados sobre o tempo;

- **Espacial:** a anomalia de dados espaciais é detectada através da comparação com os valores dos sensores vizinhos. Quando as medições são diferentes, provavelmente há necessidade de calibração. É fundamental entender se existe algum tipo de relação entre os dados quando no monitoramento multivariado, por exemplo, temperatura e umidade do ar caracteristicamente têm baixa variação espacial entre os sensores;

- **Espaço-temporal:** combinam ambas as variações espaciais e temporais, envolvendo mais sensores. Para identificação de anomalias espaço-temporal é necessário perceber que estas podem acontecer em instantes de tempo diferentes. Para isto, observam-se as interações, contabilizam-se os dados em toda a rede por certos períodos de tempo de forma local e nos sensores vizinhos.

3.2 Métodos para Detecção de Eventos em RSSF

Em RSSF, a detecção de eventos[1] pode ser utilizada em diversos cenários de aplicações. Dessa forma, várias técnicas são propostas (BAHREPOUR et al., 2009) e (PEI et al., 2014). As abordagens podem ser classificadas como (YIN; HU; YANG, 2009): baseada em limites, baseada em padrão ou baseada em aprendizagem de máquina.

Na abordagem baseada em limites a detecção ocorre quando a leitura do sensor exceder um valor limite pré-definido gerando uma notificação. As escolhas dos valores limites ficam a critério do especialista e quando há ocorrência de um evento este é informado sobre a detecção fora dos limites estabelecidos. A principal vantagem dessa abordagem é que os dados conseguem ser tratados localmente no nodo. Entretanto, o uso de um único nodo pode ser impreciso e incapaz de capturar características espaço-temporais de eventos, o que gera altas taxas de alarme no monitoramento de aplicações em redes de sensores.

A abordagem baseada em padrões detecta eventos nas leituras dos nodos através de técnicas de padrão de correspondência espaço-temporal. Um evento é detectado quando um padrão especificado pelo usuário corresponde instantaneamente a um dado recente do nodo. A principal limitação dessa abordagem é a necessidade de padrões de

[1] Sempre que não for necessário se referir explicitamente ao processo de "identificação", neste documento iremos chamar simplesmente de "detecção de eventos" o processo de detecção de *outlier* associado à identificação de um evento relevante. Da mesma forma, quando for identificado um dado espúrio iremos chamar de "detecção de dado espúrio".

eventos definidos antecipadamente.

A abordagem baseada em aprendizagem de máquina, voltada para detecção de eventos, tem sido proposta para modelar dependência espaço-temporal entre os dados dos nodos por estimação probabilística ou agrupamento. No entanto, essa abordagem assenta sobre a hipótese de independência entre as observações e a eventual indisponibilidade prévia dos dados.

A compreensão das abordagens para detecção de eventos permite escolher a mais adequada ao cenário de RSSF. De modo geral, em RSSF não é usual a disponibilidade prévia de dados, nem a localização dos nodos, os quais muitas vezes operam em ambientes hostis.

Dentre as abordagens, a baseada em limite e a baseada em padrão necessitam de informações pré-definidas para realizar a detecção. Desta forma, elas podem ser associadas à abordagem de aprendizagem supervisionada, a qual precisa do conhecimento prévio de informações para criar os rótulos para as classes. Por outro lado, a abordagem baseada em aprendizagem de máquina não necessita de informações prévias, pois por meio da probabilidade, estatísticas e agrupamentos (clusterização) esta cria seus modelos. Por tal característica, esta se relaciona à abordagem de aprendizagem de máquina não supervisionada.

Para alcançar o objetivo de estimar o alcance dos nodos foi selecionado, um método que pudesse a partir da potência do sinal recebido (RSSI), correlacionar padrões para a localização dos nodos no cenário. Cabe ressaltar que o termo alcance, neste trabalho, determina o local geográfico numa determinada região.

Este método foi selecionado por ser utilizado quando não há disponibilidade de informação e desconhecimento da posição dos nodos no cenário, características comumente presentes em RSSF de larga escala.

3.2.1 Método de Aprendizagem de Máquina não Supervisionado

A aprendizagem de máquina é o campo da Inteligência Artificial (IA) voltado ao desenvolvimento de técnicas e algoritmos destinados a ensinar a máquina, aperfeiçoando o desempenho do computador em algum processo ou tarefa.

Existem dois tipos de raciocínio na IA (RUSSELL; NORVIG, 2009):

- Indutivo: visa extrair padrões ou regras de grandes conjuntos de dados;
- Dedutivo: parte de premissas estabelecidas como verdadeiras e suas relações para deduzir novas premissas.

O interesse da aprendizagem de máquina está voltado ao raciocínio indutivo, sendo que algumas partes da aprendizagem estão intensamente relacionadas à estatística e à mineração de dados.

A aprendizagem de máquina indutiva pode ser dividida em supervisionado e não supervisionado (RUSSELL; NORVIG, 2009):

- Na aprendizagem supervisionada: é fornecido ao algoritmo de aprendizagem um conjunto de exemplos de treinamento

para os quais o rótulo da classe associada é conhecido;

- Na aprendizagem não supervisionada: o algoritmo de aprendizagem analisa os exemplos fornecidos e tenta determinar se alguns deles podem ser agrupados por semelhança ou de maneira probabilística.

Este trabalho propõe o uso da detecção de eventos baseada na abordagem de aprendizagem de máquina não supervisionada, por ser mais realística com o cenário de RSSF de larga escala, as quais não possuem informações do ambiente *a priori*. Esta aplica inferências probabilísticas, clusterização (*clustering*) e teoria dos grafos para detectar eventos.

Um aspecto importante a respeito da abordagem não supervisionada é com relação ao número de classes de separação, comumente conhecidas e definidas previamente. Contudo, em muitas aplicações essa informação não é disponível. Nesses casos, várias estratégias já foram propostas para resolução do chamado problema de validação de conjuntos (*cluster validation problem*).

O foco desta tese não é o desenvolvimento de técnica de agrupamento de dados, mas a utilização de um técnica adequada ao cenário e de RSSF. Dentre as várias técnicas, neste trabalho foi selecionada o *k-means* por suas características, a seguir discutidas.

3.2.1.1 Técnica de agrupamento *k-means*

O propósito do agrupamento de dados é separar objetos em grupos, utilizando-se de suas características inerentes. A ideia básica é que objetos similares, considerando um critério pré-estabelecido, sejam posicionados em um mesmo

grupo. O critério fundamenta-se na função de dissimilaridade, a qual retorna a distância entre os objetos. Os grupos definidos devem ser mutuamente similares e com elementos distintos para separação (ŽALIK, 2008).

No contexto da análise de dados, técnicas de agrupamento são vantajosas por reduzir o tamanho de conjuntos, simplificando a observação dos dados sobre cada grupo. Essas técnicas extraem características não triviais dos dados visando um melhor entendimento sobre a natureza do evento. Essa informação permite gerar modelos de comportamento, podendo ser utilizadas por outros métodos.

No processo de agrupamento de dados baseado em similaridade, existem diversos métodos para calcular a distância entre os dados como: distância euclidiana, distância euclidiana quadrática, distância Manhattan e distância Chebychev. Alguns métodos concentram-se em identificar os dados mais distantes do grupos. Nesse processo, o agrupamento de dados também pode identificar valores anômalos.

Neste contexto, alguns desafios são fundamentais para a clusterização e pertinentes atualmente (JAIN, 2010): (a) O que é um cluster? (b) Quais recursos devem ser usados? (c) Os dados devem ser normalizados? (d) Os dados contêm algum outliers? (e) Como definimos a similaridade pareada? (f) Quantos clusters estão presentes nos dados? (g) Qual método de agrupamento deve ser usado? (h) Os dados possuem alguma tendência de agrupamento? (i) Os clusters e partições descobertos são válidos?

Esse número de variáveis pode atribuir imprecisão na definição de um *cluster* e a dificuldade em definir uma medida

de similaridade apropriada e função objetiva.

Dentre as muitas técnicas existente para agrupamento a mais conhecida é o *k-means*. O *k-means* tem o objetivo de identificar similaridades entre os dados agrupando-os conforme um número pré-definido k de grupos. É um método simples e eficiente baseado na distância euclidiana (JAIN, 2010). Basicamente, a heurística do *k-means* busca minimizar a distância dos elementos a um conjunto de k centros (centroides) de forma interativa, sendo que a distância entre um ponto a um agrupamento é definida como sendo a distância do ponto à centroide mais próxima dele.

O algoritmo *k-means* requer três parâmetros configuração: número de *clusters* k, inicialização do *cluster* e a métrica de distância. O parâmetro mais crítico de escolha é o valor de k. Por enquanto, não existe um critério matematicamente perfeito para avaliação da escolha do k. Comumente, *k-means* é executado independentemente para diferentes valores de k e a partição que parecer ser significativa ao domínio é selecionada. Inicializações diferentes podem levar a um agrupamento final distinto por que o k significa apenas convergir para os mínimos locais. Um caminho para superar os mínimos locais é executar o algoritmo *k-means*, para um dado k, com várias partições iniciais diferentes e escolher a praticar com o menor erro quadrado. O *k-means* é tipicamente usado com a métricas euclidianas para calcular a distancia entre os pontos e o centro do *cluster*. Como resultado, o *k-means* encontra agrupamentos esféricos. O *k-means* com a distância de mahalanobis tem sido usado para detectar *cluster* hyper-elipsoidal, entretanto isso resultado em um maior

consumo computacional (JAIN, 2010).

A técnica *k-means*, representada na equação da Figura 10, é composta pelas etapas, inicialização, atribuição, movimentação e otimização (JAIN, 2010).

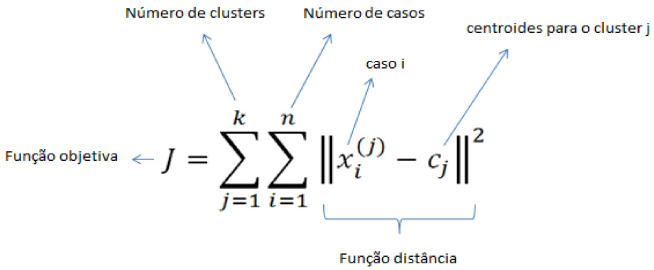


Figura 10 – Equação do *k-means*. Adaptado de: (JAIN, 2010)

1. Na etapa de inicialização, o algoritmo gera aleatoriamente k centroides, (k é um parâmetro fornecido).

2. Na etapa de atribuição são calculadas as distâncias euclidianas entre todos os pontos de dados a cada centroide. No final, baseado nas distâncias, os dados são divididos nas k centroides.

3. Após distribuir os dados nos agrupamentos, a etapa de movimentação recalcula os valores das centroides para todos os agrupamentos e atribui um novo valor médio para a centroide. O dado com localização mais próxima a esse valor médio para a centroide passa a ser a nova centroide do agrupamento.

4. A escolha de novas centroides para cada agrupamento pode fazer com que um ponto fique mais próximo de uma

centroide de outro agrupamento próximo. Na etapa de otimização, as etapas 2 e 3 se repetem até que os agrupamentos não mudem ou alcance algum critério de parada. O critério pode ser definido, por exemplo, por um número máximo de iterações.

A característica marcante deste algoritmo é a velocidade, sendo necessárias poucas iterações para convergir para o resultado, respeitando a premissa do elemento estar associado ao centro mais próximo.

Uma questão importante é quanto a homogeneidade dos dados para uma melhor separação dos grupos. Em um conjunto de dados muito homogêneo, a separação dos grupos pode ser afetada sendo necessária outra inicialização para um melhor ajuste. Outro aspecto fundamental é em relação ao número de grupos definidos na configuração do sistema. No caso de muitos grupos, o conjunto de dados pode ser dividido artificialmente. Por outro lado, um número pequeno de grupos pode ocasionar a união de conjuntos distintos de dados (ŽALIK, 2008).

O algoritmo básico do *k-means* já foi estendido com muitas finalidades. Dentre estas, aplicando heurísticas para trabalhar com tamanhos mínimos ou mesclar *clusters*. Na literatura de reconhecimento de padrões duas extensões do *k-means* são as mais conhecidas: ISODATA (BALL; HALL, 1965) e Forgy (FORGY, 1965).

No *k-means* cada ponto de dado é atribuído a um único cluster, Fuzzy c-means proposto por (DUNN, 1973) cada ponto de dados pode ser um membro múltiplo *clusters* com um

valor associado. Em (PELLEG; MOORE, 2015), o kd-tree é aplicado para identificar de forma eficiente os centros de *cluster* mais próximos para todos de dados, função principal do*k-means*. No trabalho K-medoid (GENTLE; KAUFMAN; ROUSSEUW, 2006), os *clusters* são representados usando a mediana dos dados em vez da média. Em K-Kernel (SCHÖLKOPF; SMOLA; MÜLLER, 1998), foi proposto a detectar aglomerados de forma arbitrária, com função de similaridade do *kernel*. Todas as extensões introduzem parâmetros adicionais especificados pelo usuário. Como descrito acima, muitas extensões de algoritmos foram propostos para diversas áreas de conhecimento, aumentando a complexidade da revisão para todas as abordagens publicadas.

Com o surgimento de novas aplicações, tornou-se cada vez mais claro que a tarefa de buscar o melhor princípio de agrupamento poderia ser inútil. Um método de agrupamento que satisfaça os requisitos para um grupo de usuários pode não satisfazer os requisitos de outro (JAIN, 2010). Por fim, agrupamento de dados deve envolver as necessidades do usuário ou do aplicativo para melhorar compreensão dos resultados gerados.

3.3 Métodos para Detecção de Outliers em RSSF

Obter informações úteis ao processo decisório a partir de dados bruto, levando-se em conta a natureza complexa e dinâmica das RSSF é o desafio dos novos mecanismos de detecção de *outliers*. Tal desafio está na eficácia dos métodos de detecção com respeito à precisão de detecção, taxa de detecção e alarmes falsos. Também com relação à eficiência na utilização

dos recursos há os desafios relacionados ao consumo de energia, memória, processamento e transmissão.

Os trabalhos de Zhang, Meratnia e Havinga (2010), Rassam, Zainal e Maarof (2013a) apontam alguns desafios ao projeto e adaptação para soluções em RSSF:

- **Limitações de recursos computacionais:** os nodos de baixo custo têm limitação de memória, processamento, consumo de energia e taxa de transmissão dos dados. No processo de detecção, os recursos mais exigidos para análise das anomalias são processamento e armazenamento. O consumo de energia é um dos principais desafios em RSSF, onde o gasto energético nas atividades de transmissão dos dados são milhares de vezes maiores do que o custo de processamento (HILL et al., 2000). Portanto, o desafio na detecção de *outlier* está na forma de minimizar o consumo energético e simultaneamente equilibrar a utilização dos recursos computacionais;

- **Sobrecarga na comunicação:** A maioria das técnicas de detecção de anomalias utiliza a abordagem centralizada na análise, onde os dados são coletados pelos nodos em enviados para a estação base. Um desafio é reduzir a transmissão dos dados para prolongar a vida útil dos nodos, criando novos modelos para detecção de forma distribuída;

- **Topologia de rede dinâmica e heterogênea:** o dinamismo das aplicações associadas à mobilidade dos nodos pode gerar aumento de falhas na comunicação.

Tais mudanças influenciam negativamente na validade do modelo de referência utilizado para detectar as anomalias. Outro aspecto da heterogeneidade aparece quando os dados coletados por nodos são analisados por um modelo de detecção de anomalias que utiliza um modelo de referência de outra rede. Essa dinamicidade e heterogeneidade aumenta a complexidade de projetar técnicas de detecção de anomalias.

- **Escalabilidade de rede:** a quantidade de nodos da rede de sensores pode alcançar milhares de nodos sensores. Como principal desafio, as técnicas de detecção de *outliers* devem manter uma alta taxa de detecção e baixa taxa de falsos alarmes. Modelos de referência com a idade dos sensores devem ser atualizados, sendo uma tarefa muito difícil para aplicações de grande escala. Além disso, existe o desafio das técnicas de *outliers* terem dificuldades em expandir a quantidade do fluxo de dados de forma distribuída e *online*;

- **Aumento da dimensão dos dados:** a dimensão dos dados pode aumentar, como consequência, o consumo computacional do sensor também;

- **Identificação da fonte de *outlier*:** os dados brutos são fornecidos por sensores por acontecimentos ocorridos na rede. Entretanto, há dificuldades na identificação da causa do *outlier* detectado pela limitação de recursos dos nodos e pela dinâmica da RSSF;

- **Distinção entre tipos de outliers:** As técnicas de detecção de *outlier* muitas vezes não fazem distinção

entre dados espúrios e eventos relevantes, tratando todos como dados espúrios. Dessa forma, resultando em perda de informações ocultas importantes sobre eventos. Assim, um desafio de detecção de anomalias em RSSF é como identificar fontes *outlier* e fazer tratamento entre dados espúrios e eventos. Por fim, o grande desafio das técnicas de detecção de *outlier* é balanceamento entre os requisitos de precisão e o consumo dos recursos.

3.3.1 Requisitos para Classificação dos Métodos de Detecção de *Outliers*

Nesta seção, apresentam-se os requisitos para classificar os métodos de detecção de *outliers* em RSSF. Estes requisitos são utilizados como parâmetros para comparar características de diferentes métodos de detecção.

Os requisitos considerados como métricas para (RASSAM; ZAINAL; MAAROF, 2013a), (ZHANG; MERATNIA; HAVINGA, 2010) e (BHOJANNAWAR; BULLA; DANAWADE, 2013) são:

- Entrada de dados;
- Identificação de *outliers*;
- Modelo de detecção;
- Modo de operação;
- Grau do *outlier*;
- Adaptabilidade a mudanças;
- Disponibilidade de dados pré-definidos.

Entrada de dados: A identificação do tipo do dado determina qual técnica de detecção é mais adequada para sua análise. Por outro lado, o grande volume de dados pode sobrecarregar a transmissão aumentando o consumo de energia do sensor. Portanto, adotar algumas estratégias para reduzir a quantidade de dados que são transmitidos e também minimizar a complexidade em termos de comunicação e computação são desejáveis. São considerados dois aspectos para classificação do tipo de dado: atributos e correlação.

Atributos: classificando dados com base na suas dimensões, podendo ser univariado ou multivariado. A transmissão de dados multivariados pode sobrecarregar a rede e aumentar o consumo energético. Portanto, a redução da dimensão dos dados multivariados pode prolongar a vida no nodo através da diminuição da utilização do rádio;

Correlações: existem dois tipos de dependências em nodos: i) as dependências entre atributos no nodo. ii) as dependências entre leituras no nodo e leituras entre vizinhos. Atributos multivariados podem correlacionar dados. Por exemplo, a leitura da umidade e da pressão estão relacionadas com a leitura da temperatura. A captura das correlações de atributos ajuda a melhorar a precisão da mineração dos dados e a eficiência computacional. Nesse contexto, os dados dos sensores podem ser correlacionados em tempo e espaço, em especial para o monitoramento de ambientes. A correlação temporal implica na existência de observações no instante de tempo e relaciona a outras leituras em instantes de tempo anteriores. Enquanto a correlação espacial implica que as leituras dos nodos geograficamente próximos estão

correlacionadas uns com os outros em grande parte. A captura da correlação espaço-temporal ajuda a prever a estimativa das leituras e distinguir dados espúrios de eventos relevantes em RSSF.

Identificação do *outlier*: existem três fontes de *outliers* em RSSF: ruído e erros, eventos e ataques maliciosos.

Erros: refere-se a uma medição relacionada com o ruído ou dados provenientes de um sensor defeituoso. A ocorrência de erros com frequência tem, como consequência, a sua identificação como dado espúrio. A qualidade dos dados é influenciado pelo número de *outliers*, que precisam ser identificadas e corrigidas, posteriormente ainda passar por uma nova análise;

Eventos: um evento pode ser definido como um fenômeno especial que altera o estado do mundo real, por exemplo, incêndio florestal. Este tipo de anomalia muda o padrão histórico dos dados do sensor e tem duração relativamente de longo tempo. Entretanto, a similaridade de dados anômalos gerados entre sensores defeituosos e um evento é muito tênue. Para distinção é necessário fazer uso de dados de nodos vizinhos e de similaridade espaço-temporais dos dados dos sensores. Este fato se sustenta devido às susceptibilidade das falhas serem espacialmente não relacionadas, enquanto as medições de eventos são susceptíveis de serem correlacionadas espacialmente (KRISHNAMACHARI; IYENGAR, 2003) (HUTCHISON; MITCHELL, 2004);

Ataques maliciosos: redes de sensores podem ser utilizadas para fins de segurança ou aplicações militares. Neste contextos estão sujeitas a manipulações não desejadas. Este

tipo de anomalia é tratada com técnicas de detecção de intrusão comumente encontradas em trabalhos de tolerância a falhas bizantinas.

Modelo de Detecção: as técnicas de detecção existentes utilizam os seguintes modelos de estrutura: local, centralizado e distribuído.

Local: a identificação dos dados anômalos ocorre nos nodos individuais. A detecção de *outliers* local tem como vantagens a diminuição da transmissão de dados e a melhoria na escalabilidade. Existem duas formas de identificação de *outliers* no modelo local. Na primeira, os nodos utilizam os históricos dos seus valores para identificação de *outliers*. A outra forma, além do histórico das leituras, também utiliza leituras dos nodos vizinhos para identificar, de forma colaborativa, as anomalias. Contudo, analisando as duas formas, a segunda oferece melhor precisão e robustez na detecção utilizando correlação espaço-temporal dos nodos;

Centralizada: no modelo centralizado, todos os dados são enviados para uma localização central, como a estação base ou *cluster-head*, onde acontece o processo de detecção de anomalias. Este mecanismo sobrecarrega a rede de comunicação e retarda o tempo de resposta.

Distribuído: esse modelo adota a colaboração entre os nodos para detecção, no qual cada nodo envia um resumo de seus dados representados por um modelo de referência local para estação base. A partir dos vários modelos de referência locais é construído o modelo de referência global. O modelo de referência global é enviado para todos os nodos para ser

utilizado com parâmetro na detecção de anomalias. Embora este modelo distribuído otimize o tempo de resposta e consumo de energia, por outro lado, existe o problema da abordagem centralizada para redes de sensores extensas.

Modo de operação: quanto ao modo de operação em RSSF este pode ser: *online* ou *offline*.

No modo *online* a detecção de anomalias é realizada imediatamente após a leitura do sensor. A detecção *online* é preferível para minimizar os tempos de atraso e assegurar a integridade dos dados. Entretanto, o custo do método de detecção é alto em termos de consumo de recursos (energia, processamento, armazenamento e largura de banda). A eficiência de soluções de detecção *online* está diretamente relacionada à complexidade computacional dos métodos utilizados;

No modo *offline* a detecção de anomalias é realizada após uma janela de tempo específica. Embora tenha menor consumo de energia, requer mais memória para o armazenamento de lotes de dados por um período de tempo. Por outro lado, a integridade dos dados pode ser afetada pelo tempo de atraso na detecção.

Grau do *Outlier*: grau de mensuração de dados que desviam do padrão normal. As técnicas de detecção de anomalias podem não somente identificar o *outlier* mas também fornecer métodos específicos para calcular o grau de desvio. Em RSSF, as anomalias são medidas em duas formas: escalar e pontuação do *outlier*.

Escalar: na avaliação escalar, a medição classifica um dado como normal ou *outlier* atribuindo valor zero ou um. As técnicas que fazem uso desse requisito não diferenciam valores extremos e nem fornecem lista ordenada de *outliers*, apenas apresentam lista de medidas normais e outra das anomalias.

Pontuação do *outlier*: nessas técnicas, há atribuição de pontuação ao *outliers* para cada medição realizada. Uma análise classifica as anomalias em altas ou baixas por pontuação. Atribuir um valor limite para classificação exige conhecimento sobre a grandeza monitorada e do cenário. A solução ideal para RSSF é aprender o limite e atualizá-lo constantemente.

Adaptabilidade a mudanças: o modelo que representa o comportamento normal dos dados deve ser atualizado para uma eficaz detecção de anomalias. Isso devido a dinâmica na transmissão dos dados medidos pelo sensor. Por outro lado, o mecanismo de adaptação das mudanças deve ser eficiente e leve para não consumir os limitados recursos do nodo.

Disponibilidade de dados pré-definidos: uma maneira simples para identificar anomalias é a construção de um modelo de referência para dados normais e outro para detecção de *outlier*. A leitura que diverge significativamente deste modelo de referência normal é classificado como *outlier*. Quanto há disponibilidade do modelo de referência, podemos classificar as técnicas de detecção de *outliers* em três categorias básicas: abordagem supervisionada, não supervisionada e semi supervisionada. Na abordagem supervisionada e semi

supervisionada, um modelo de referência pré-existente para classificar dados em normais ou anômalos é exigido. Em aplicações de RSSF não é usual a disponibilidade de um modelo de referência para classificação dos dados. Por outro lado, na abordagem não supervisionada não há necessidade da pré-existência do modelo de referência, pois constrói o próprio através de estimações.

3.3.2 Classificação dos Métodos de Detecção de *Outliers*

Nesta seção, apresenta-se a classificação dos métodos de detecção de *outliers* para RSSF, baseadas em (RASSAM; ZAINAL; MAAROF, 2013a), (ZHANG; MERATNIA; HAVINGA, 2010),(BHOJANNAWAR; BULLA; DANAWADE, 2013): estatística, vizinhança, clusterização, classificação e decomposição espectral. Estas abordagens se diferenciam quanto à técnica utilizada para detecção, consumo dos recursos, especificidade da aplicação e cenário.

Detecção de *Outliers* baseada em Estatística

As técnicas de detecção baseada em estatística trabalham com modelos, assumindo conhecimento prévio ou estimando (probabilidade ou grafos) a distribuição dos dados e avaliando-os na forma como se encaixam no modelo referência. É considerado como anomalia qualquer dado cujo valor se afasta do modelo de referência. Com relação ao conhecimento sobre a distribuição dos dados, a abordagem estatística é classificada em paramétrica e não paramétrica.

Na categoria paramétrica é assumida a disponibilidade de conhecimento sobre os dados. Baseada no tipo de

distribuição assumida na categoria paramétrica, esta é dividida em gaussiana e não gaussiana. No modelo gaussiano, a distribuição dos dados assume o comportamento de uma distribuição normal. No modelo não gaussiano, outras funções de densidade de probabilidade são utilizadas.

Na categoria não paramétrica, a distribuição de dados não é conhecida antecipadamente. Neste contexto, as abordagens baseadas em histograma e estimador densidade *kernel* são amplamente utilizadas.

O modelo de histograma envolve a contagem da frequência de ocorrência de diferentes dados (estima a probabilidade de novas ocorrências). Na sequência, compara a ocorrência desta estimação com todas as categorias do histograma e verifica se pertence a uma delas.

No modelo de estimador de densidade *kernel* são usadas funções do *kernel* para estimar a probabilidade da função distribuída (PDF) para casos normais. Um novo caso, que reside na área de baixa probabilidade é declarado como uma anomalia.

Contudo, a abordagem baseada em estatística pode ser eficiente na identificação de *outliers* se o modelo de distribuição probabilística for obtido. Em cenários comuns de RSSF, não há disponibilidade do conhecimento sobre a transmissão de dados dos sensores. Então a categoria paramétrica pode ser inútil se os dados dos sensores não seguirem a distribuição normal. As técnicas não paramétricas são atraentes devido ao fato que não fazem qualquer suposição sobre as características da distribuição. O modelo histograma é muito eficiente para dados univariado, por outro lado, não são capazes de observar

relações de diferentes atributos (e.g. temperatura e umidade) para dados multivariados. A função *kernel* pode ser escalável para dados multivariados e consumir poucos recursos computacionais.

Detecção de *Outliers* baseada em Vizinhança

As técnicas baseadas nesta abordagem utilizam a detecção de anomalias em rede, com a suposição de que os padrões normais dos dados são sempre encontrados pela vizinhança, enquanto os anômalos estão longe de seus vizinhos (CHANDOLA; BANERJEE; KUMAR, 2009). A premissa deste modelo é a utilização de medidas de similaridade, por exemplo, medida de distância Euclidiana ou distância de Mahalanobis, que medem o grau do padrão normal dos dados ou do padrão da anomalia.

A limitação da abordagem baseada em vizinhança está no consumo computacional para o cálculo da distância entre padrão de dados em conjunto de dados multivariado. Como consequência, o modelo não possui boa escalabilidade.

Detecção de *Outliers* baseada em Clusterização

Essa abordagem de clusterização (*clustering*) propõe o agrupamento de padrões de similaridade entre os agrupamentos de dados (*clusters*) através de modelos de mineração de dados. Um agrupamento é considerado anômalo se for distante de outros agrupamento do conjunto de dados (RASSAM; ZAINAL; MAAROF, 2013b). Medidas de similaridade são utilizadas para determinar a afiliação do padrão de dados em um agrupamento. Assim, a distância Euclidiana pode ser

utilizada. Técnicas de clusterização vêm sendo utilizadas em redes de sensores para encontrar caminhos mais eficientes para comunicação e processamento de dados.

A utilização da abordagem de clusterização ajuda a reduzir o custo de comunicação, agrupando os dados localmente. Não necessitando também de conhecimento prévio dos dados, sendo totalmente não supervisionado. As principais desvantagens identificadas na utilização das técnicas de clusterização apresentadas em (RASSAM; ZAINAL; MAAROF, 2013a) são:

- Essa técnica é computacionalmente dispendiosa com dados multivariados, porque os cálculos das medidas das distâncias entre todos os padrões de dados têm alto custo computacional, tornando-se inviáveis para uso em nodos com limitados recursos computacionais;

- A dependência sobre a escolha da largura de banda do agrupamento em algumas situações as tornam inadequadas para aplicações de RSSF;

- Essas técnicas possuem dificuldades para tratar mudanças contínuas do fluxo dos dados ao longo do tempo.

Contudo, a atualização do modelo de referência é fundamental no momento de sua utilização. Alguns modelos recentes como (MOSHTAGHI et al., 2012) utilizam o aprendizado incremental para melhorar a atualização, por outro lado, o custo computacional é muito alto para a natureza limitada da RSSF.

Detecção de *Outliers* baseada em Classificação

Essa abordagem faz uso de importantes modelos de aprendizado de máquina e mineração de dados em que um classificador é treinado usando padrões de dados conhecidos para aprender o modelo de classificação. Classificadores não supervisionados com múltiplos modelos de classes não são utilizados em RSSF, devido a dificuldade em classificar os sensores. Contudo, classificadores não supervisionados com apenas um modelo de classe são mais adequados para detecção de *outlier* em RSSF, por definirem apenas um padrão normal, os dados fora deste limite são considerados anomalias.

Os classificadores precisam atualizar o modelo de classificação padrão para se ajustar aos novos conjuntos de dados. As técnicas de detecção de *outliers* para RSSF baseadas na abordagem de classificação são: máquina de vetor de suporte (SVM) e redes bayesianas.

Máquina de vetor de suporte: a técnica separa as classes diferentes de dados ajustando um hiperplano entre eles que minimiza a separação. Os dados são mapeados num espaço de características de dimensão superior, onde pode ser facilmente separado por um hiperplano (conjunto de dados que pode ser separado por um plano). Além disso, uma função *kernel* é usada para aproximar os produtos do ponto entre os vetores mapeados no espaço característico para encontrar o hiperplano;

Redes bayesianas: essa categoria usa um modelo gráfico probabilístico para representar um conjunto de variáveis e suas independências probabilísticas. Essas redes agregam informações de diferentes variáveis e proporcionam uma estimativa sobre a expectativa do evento que pertence à classe

aprendida. Existe uma divisão nesta categoria quanto ao grau de independência entre as variáveis em rede bayesiana:

- As técnicas bayesianas ingênuas: capturam correlações espaço-temporais entre os nodos;
- As técnicas bayesianas crenças: consideram as correlações entre os atributos dos dados do nodo;
- As técnicas bayesianas dinâmicas: consideram a topologia da rede e sua dinâmica ao longo do tempo, adicionando novas variáveis de estado para representar o estado do sistema no instante de tempo atual.

A abordagem de classificação fornece um conjunto exato de anomalias através da construção do modelo de classificação. Entretanto, uma desvantagem da técnica baseada em Máquina de Vetor de Suporte (SVM) é sua complexidade computacional e a escolha da função kernel apropriada. Igualmente, redes bayesianas têm dificuldade na exatidão do modelo de classificação, se o número de variáveis for grande para a RSSF.

Detecção de *Outliers* baseada em Decomposição Espectral

Essa abordagem visa encontrar padrões de comportamento nos dados usando análise de componentes principais. A análise de componentes principais (PCA) é uma técnica que é utilizada para reduzir a dimensionalidade antes da detecção do *outlier* para encontrar um novo subconjunto de dimensão que capture o comportamento dos dados. Notadamente, poucos PCA capturam a construção da

dimensionalidade e qualquer dado que viole essa estrutura por menor que seja é considerado como um *outlier*.

Um desafio dessa abordagem está na seleção dos componentes adequados, necessários para estimar com exatidão a matriz de correlação dos padrões normais, que consumem muitos recurso computacionais.

3.3.3 Descrição das Técnicas Leves Baseadas em Estatística para Detecção de *Outliers*

Na literatura existem vários trabalhos relacionados à taxonomia para técnicas de detecção de *outliers*, como (MARKOU; SINGH, 2003), (MARKOS; SINGH, 2003), (HODGE; AUSTIN, 2004) e (CHANDOLA; BANERJEE; KUMAR, 2009). Em geral, a classificação destas técnicas ocorre a partir de diferentes perspectivas.

Neste trabalho foi utilizada a abordagem taxonômica desenvolvida especificamente para RSSF, na qual as técnicas são classificadas por (ZHANG; MERATNIA; HAVINGA, 2010), detalhes na Seção 3.3.2.

Cada uma das abordagens possui especificidades e, portanto, pode ser melhor adequada para determinados cenários de aplicação. Exemplificando, nas abordagens baseadas em vizinhos, em classificação e em decomposição, a forma utilizada para detecção de *outliers* tem custo computacional alto por serem fundamentadas em mineração de dados e aprendizado de máquina. Por outro lado, na abordagem baseada em clusterização, o desafio está na escalabilidade de rede e no fluxo dos dados (ZHANG;

MERATNIA; HAVINGA, 2010).

No estudo realizado nesta tese foram consideradas apenas abordagens estatísticas e paramétricas *stateless*. Nesse sentido, as informações sobre o histórico de dados a partir de cada sensor não foi considerado na detecção. Os métodos que foram considerados nesta pesquisa são discutidos abaixo.

Método Peirce

Critério do Peirce é um método estatístico para a detecção de *outlier* baseado em uma distribuição normal. De acordo com Ross e Ph (2003), o método de Peirce foi descrito em 1852 como: "as observações devem ser rejeitadas quando os desvios reais da média obtidos por mantê-los é menor do que os desvios obtidos por sua rejeição, multiplicado pela probabilidade de fazer tantos e não mais, observações anormais".

Resumidamente, o principal objetivo do Critério de Peirce é gerar probabilidades de erro no sistema, assumindo n amostras e k valores suspeitos (medições duvidosas). O método começa com apenas um k valor suspeito e compara a média obtida com a distribuição normal para decidir se rejeitará ou não o valor suspeito. A cada rodada, o método incrementa k (o número de medições duvidosas) até que não seja mais necessário eliminar dados, permitindo a detecção de mais do que um *outliers* simultaneamente. Apesar de ser proposto há mais de 150 anos, o método de Peirce é usado ainda hoje. Como o método proposto original tem um certo custo computacional, geralmente utiliza-se uma tabela própria para simplificar os cálculos.

A aplicação do método Pierce é apresentada na Figura 11.

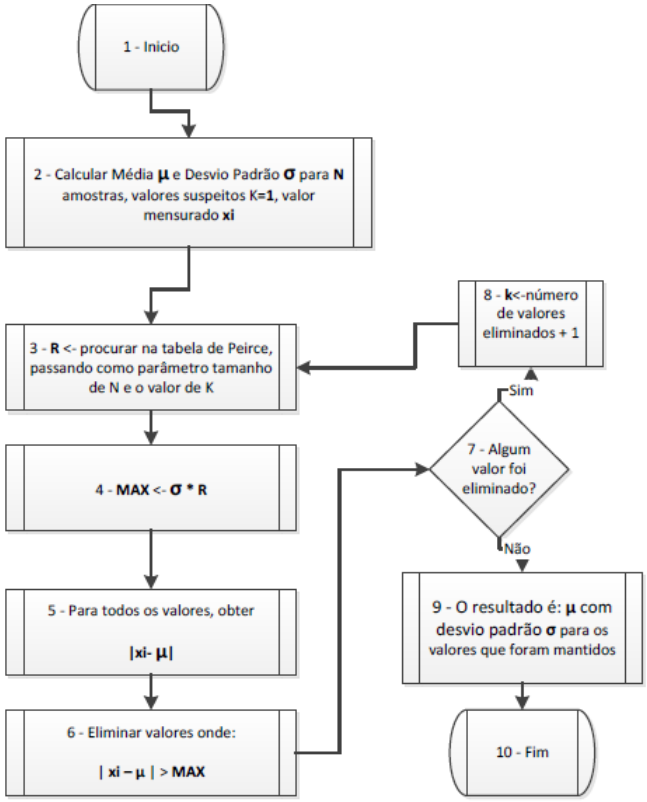


Figura 11 – Fluxo do método de Pierce. Adaptado por: (CALLEGARO, 2014). Fonte: (ROSS; PH, 2003)

Técnica de Chauvenet

Esse também é um método estatístico para a detecção de anomalias em uma distribuição. O critério se baseia na

hipótese de que uma medida arbitrária pode ser rejeitada se a probabilidade de obter o desvio da média para este valor é menor do que o inverso do dobro do número de medições (TAYLOR, 2012).

Para utilização deste método é fundamental observar o tamanho da amostra. Para grandes amostras, a média é afetada de forma insignificante, ao menos que o valor observado esteja muito longe da média para movimentar a distribuição.

No critério de Chauvenet aplica-se a média ($\bar{x}$) e o desvio padrão (s) do conjunto de dados. O desvio de cada um dos valores em relação à média é comparado ao desvio padrão, multiplicado pela constante de Chauvenet (k) conforme equação 3.1.

$$|x_i - \bar{x}| > k_{(n)}.s \tag{3.1}$$

A condição de rejeição será a diferença de valores em relação a média, sendo maior ao valor (k) multiplicado pelo desvio padrão. A tabela 1 mostra a relação do número de observações (n) em relação os valores de (k).

Tabela 1 – Valores de k.

n	**k(n)**
2	1,15
4	1,54
6	1,73
8	1,86
10	1,96

Média Tolerante a Falhas - MTF

Método bem simples proposto por Marzullo (1990), para fusão de sensores. O método ordena e divide em três partes um conjunto de dados, excluindo os extremos. A ideia básica, t=N/3, onde N é o tamanho do conjunto, exclui os extremos da amostra ordenada, ao final, calcula o desvio padrão e da média do valores remanescentes.

Método de Ponderamento por Confiança - CWA

O método proposto por Elmenreich (2007), está relacionado com a confiança nos sensores baseado no valor da variância nos seus dados lidos. A utilização da variância mede a dispersão dos dados, podendo comparar diferentes grupos. De forma conceitual, mede quão distante os valores estão da média, isto é, a confiança nos dados do sensor é inversamente proporcional a sua variância.

Uma vez que o Método Ponderamento por Confiança (CWA) usa a variância de cada sensor como uma métrica para calcular o peso associado com o sensor, este prioriza sensores com variância menor. Problemas são detectados quando há mau funcionamento do sensor e envio do mesmo valor repetidamente. Esta situação gera valores de variância muito perto de zero. Neste sentido, os sensores com valores de variância iguais ou muito próximos de zero são assumidos como defeituosos, e suas variância são artificialmente aumentadas para um valor limite do CWA. Importante notar CWA não detecta casos anômalos. No entanto, na prática, é possível afirmar que os valores extremos são removidos ou, pelo menos, “parcialmente removidos” pois têm reduzidos os seus pesos no cálculo da média.

Para o cálculo da média ponderada, assume-se que as observações são recebidas da mesma entidade, e o domínio de valor é contínuo. As observações devem ser feitas aproximadamente no mesmo instante. As funções de erro são consideradas independentes, os valores de medição são fundidos usando-se uma média ponderada conforme a equação 3.2.

$$P = \frac{\sum_{i=1}^{n} S[i] \frac{1}{V[i]}}{\sum_{i=\frac{1}{V[i]}}^{n}} \tag{3.2}$$

Onde n é o número de observações de entrada, Si representa os valores da medição e Vi é a variância estimada.

Método CWA+MTF

Elmenreich (2007) afirma que o valor médio calculado pelo método CWA pode ser concebido como um valor de referência. Além disso, segundo o autor, o método CWA pode ser melhorado associando seus resultados com outros métodos de tratamento de *outliers*.

Neste sentido, a associação de CWA com o método de Média Tolerante a Falhas (MTF) torna possível ponderar o valor obtido usando CWA e efetuar uma subsequente detecção e remoção de valores extremos através do MTF. Especificamente, a proposta CWA + MTF consiste em calcular a média ponderada de acordo com o método CWA e, posteriormente, remover 2/3 dos dados de acordo com o método MTF. No final, a média é calculada a partir dos dados restantes.

$$S1[i] = \frac{P - S[i]}{V[i]} \tag{3.3}$$

Onde Vi é a variância do sensor Si e aplicar a MTF. Os valores restantes formam o conjunto de sensores e suas variâncias, da qual deverá ser novamente calculada a média ponderada.

3.4 Diferenças entre Métodos de Detecção de *Outliers* e Detecção de Eventos

Uma vez identificado o tipo de fonte da anomalia, é necessário fazer a correta identificação do *outlier*. Dessa forma, se o *outlier* é detectado como um erro ou dado ruidoso (dado espúrio), deve ser removido a partir dos dados detectados para conferir qualidade na precisão dos dados e economia no uso da bateria através da eliminação de dados na transmissão. Entretanto, se o *outlier* for gerado por um evento relevante (e.g. incêndio), a eliminação do *outlier* levará à perda da informação sobre o acontecimento, que pode ser relevante para entendimento da causa inicial.

As principais diferenças entre a detecção de *outliers* e a detecções de eventos envolvem (ZHANG; MERATNIA; HAVINGA, 2010):

- Técnicas de detecção de *outliers* não têm conhecimento *a priori* da condição de disparo ou semântica de qualquer evento; enquanto as técnicas de detecção de evento mantêm a condição de disparo ou semântica de determinado evento enviado pelo nodo central;

- Detecção de *outlier* visa identificar leituras anômalas comparando as medições de sensores uns com os outros, enquanto a detecção de evento tem por finalidade estabelecer a existência de um determinado evento comparando as medições do sensor com a condição de disparo ou padrão pré-definido;

- A técnica de detecção de *outliers* precisa impedir que o dado normal passe a ser classificado como *outlier*, assim mantendo alta taxa de detecção e baixa taxa de alarme falso. Enquanto a técnica de detecção de evento precisa impedir que dados espúrios que estão em conformidade com a condição de evento influencie na confiabilidade da detecção;

Por outro lado, a característica em comum nas técnicas de detecção de eventos é a utilização de correlação espaço-temporal entre dados do nodo e dos seus vizinhos para distinguir entre eventos e dados espúrios. Esta se baseia no fato de que as medições ruidosas e falhas dos sensores são aleatórias e não relacionadas; enquanto as medições de eventos são susceptíveis de ser correlacionado espacialmente (KRISHNAMACHARI; IYENGAR, 2003).

Devido ao fato de nem todas as anomalias serem identificadas em aplicações de detecção de eventos, as técnicas de detecção de *outlier* podem ser adaptadas para o domínio da detecção de eventos. A escolha sobre a técnica de detecção passa pela compreensão da fonte e tipos de anomalia para seu devido tratamento.

3.5 Métodos de localização de nodos baseado na Estimação da Conectividade

A definição dos critérios para relação da conectividade entre os nodos objetivando a localização destes tem várias vertentes. Nesse sentido algumas hipóteses foram levantadas para estimação da auto localização dos nodos (CHENG et al., 2012a), (YOUSSEF; YOUSSEF, 2007):

- Método baseado na potência do sinal recebido;
- Método baseado no tempo de chegadas das mensagens;
- Método baseado no ângulo de chegada do sinal recebido;
- Método baseado na diferença dos tempos de chegada das mensagens.
- Método de correspondência entre padrões;
- Método de baseado na contagem de *hops*.

Entre essas alternativas, o método de potência de sinal recebido foi o escolhido nesta tese por possuir características desejáveis ao cenário de RSSF, os detalhes desse método são descritos a seguir.

A primeira é com relação ao conhecimento da localização dos nodos no cenário. Existem algumas propostas de taxonomias para a estimação da localização de nodos (CHENG et al., 2012a), (YOUSSEF; YOUSSEF, 2007). Estas abordam o mecanismo de estimação em duas classes:

- Baseado em âncora: operam com a necessidade de nodos âncoras, os quais têm sua posição conhecida de forma manual ou por GPS. O objetivo neste caso é através do conhecimento dos nodos âncoras estimar a posição dos nodos desconhecidos;

- Livre âncora: não fazem qualquer suposição em relação à posição do nodo. Neste caso a estimação da posição é relativa, na qual o sistema de coordenadas é estabelecido por um grupo nodos de referência.

A implantação de nodos em RSSF de larga escala com conhecimento da posição tem altos custos financeiros e operacionais. Deste modo, a seleção pelo livre âncora neste trabalho foi por essa classe se adequar ao cenário usual de RSSF, em especial a grandes áreas.

Partindo da abordagem livre âncora, há duas categorias em relação a forma de estimar o alcance dos nodos:

- Baseada em alcance: utiliza o alcance absoluto ponto-a-ponto para estimar a localização. Dentre os métodos desta categoria destacam-se: Tempos de Chegada, Diferença entre os Tempos de Chegada, Ângulo de Chegada e o Indicador da Potência de Sinal Recebido.

- Livre alcance: não faz qualquer suposição sobre a disponibilidade ou validade de tal informação. Nessa categoria destacam-se: método de correspondência de padrão e método baseado em contagem de *hop*.

O método selecionado para estimação foi o baseado em alcance, por estimar através da conectividade da rede alcances dos nodos por meio do RSSF de forma ponto-a-ponto. Também por retratarem mais adequadamente o cenário de RSSF.

O método estudado nesta tese para estimação foi o baseado em alcance, em especial, potência do sinal recebido. Esta escolha é fundamentada no custo-benefício e pela adaptabilidade ao cenário de RSSF.

.

3.6 Análise dos Trabalhos Relacionados

Nesta seção são discutidos os principais trabalhos relacionados à tese. Por ser tratar de uma arquitetura multicamada, esta pesquisa foi elaborada em duas linhas principais: detecção de eventos em RSSF de larga escala e estratégia de formação de *cluster-tree* baseada em dados;

Existem várias abordagens para detecção *outliers* (seção 3.3.2). Estas possuem suas especificidades e a escolha da abordagem adequada está diretamente relacionadas aos cenários de aplicação, Figura 9. O processo de detecção é a primeira etapa essencial na detecção de *outliers*. Neste trabalho de doutorado foi assumido o uso de abordagem para detecção de *outliers* baseada em estatística, principalmente devido a sua adequação ao cenário de RSSF de larga escala e por requer pouca capacidade computacional dos nodos, Seção 3.6.1.

Quanto à formação da rede, foi selecionada a topologia *cluster-tree* por ser mais eficiente com relação a atrasos de comunicação e consumo energético em RSSF. A partir da

topologia selecionada, foi desenvolvida a estratégia de formação baseada em dados para redes *cluster-tree*. Na Seção 3.6.2 são discutido e categorizados os trabalhos relacionados com a estratégia proposta.

3.6.1 Trabalhos Correlatos - Detecção de Eventos

Os trabalhos abaixo discutem abordagens para detecção de eventos relacionadas com o tema desta tese. A Tabela 2 faz a comparação entre trabalhos que possuem abordagens baseadas em estatística para detecção de *outliers*.

Pei, 2014

O trabalho de Pei et al. (2014) propõe uma arquitetura distribuída composta por camadas para detecção de eventos. Esta estrutura hierárquica estabelece uma sequência de ações quando da detecção do evento, representada na Figura 12.

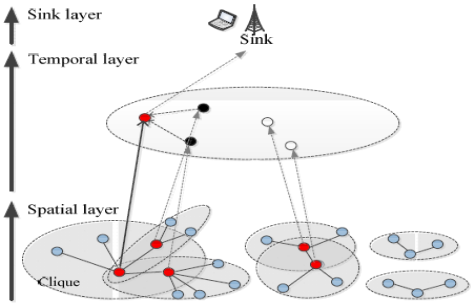


Figura 12 – Modelo Hierárquico para Detecção de Eventos. Fonte: (PEI et al., 2014)

A camada inicial é a espacial, a qual tem a função de processamento da detecção de eventos usando nodos e aplicando teorias de Campos aleatórios de Markov. Nesta camada, cada nodo comunica-se com seus vizinhos mais próximos inferindo sobre avaliação do evento. A comunicação é definida pelo alcance do rádio aplicando a equação da distância euclidiana para determinar a proximidade do nodo em relação aos seus vizinhos. Uma vez conhecidos os vizinhos, suas leituras são fundidas e posteriormente enviadas à camada temporal. O propósito é reduzir a quantidade de transmissão de dados, economizar energia e aumentando a precisão na detecção. A camada temporal faz a ligação entre a camada espacial e a estação base, com a função de processamento da detecção de eventos sobre o período de tempo, aplicando cadeia de Markov. A função da cadeia de Markov é modelar o relacionamento das leituras dos dados recebidos pela camada espacial. Os eventos marcados espaço-temporalmente são encaminhados para a estação base.

Mousavi, 2013

O trabalho de Mousavi et al. (2013) propõe um modelo descentralizado para detecção de eventos espaço-temporais para rede de sensores sem fio. A proposta explora as dependências temporais e espaciais dos nodos no ambiente distribuído. Para correlação espacial, a proposta aplica o conceito de campos aleatórios de Markov na geração de grafos, a partir da probabilidade particular do estado observado em detrimento dos demais estados de outros nodos. Quanto à questão temporal, a teoria de cadeia de Markov para modelar

as transições das leituras dos nodos no tempo é aplicada. Os resultados apresentados da abordagem demonstram um melhor desempenho quando comparado à abordagem baseada em limites, em termos de precisão e escalabilidade comparativamente com a abordagem centralizada.

Oladimeji, 2015

O trabalho de Oladimeji, Smiee e Mieee (2015) propõe uma abordagem híbrida para detecção de eventos (incêndios) em ambientes. Esta proposta, inicialmente faz a identificação do evento através da técnica *k-means* de clusterização criando grupos, conforme a Figura 13.

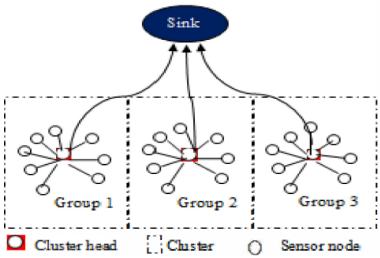


Figura 13 – Modelo de Redes de Sensores baseado em *Clusters*. Fonte:(OLADIMEJI; SMIEE; MIEEE, 2015)

Após a geração dos grupos são aplicadas técnicas de aprendizagem de máquina supervisionada, com padrões previamente estabelecidos para detecção de anomalias. O trabalho apresenta bons resultados na taxa de detecção de eventos nos testes de desempenho. Entretanto, o consumo dos recursos computacionais é alto.

Bettencourt, 2007

No trabalho de Bettencourt (2007) foi desenvolvido um algoritmo para detecção de eventos com a possibilidade de operar de forma *online* para estimar eventos em tempo real. Para tanto, as leituras anteriores dos nodos devem ser conhecidas e cada nodo aprende a distribuição estatística das diferenças entre a sua medição e a dos seus vizinhos. Para modelar as transições das leituras dos nodos no tempo utiliza-se cadeia de Markov e teste de significância para os dados da distribuição.

Comparação da proposta desta tese com outros trabalhos

Esta tese de doutorado destaca os trabalhos relacionados às arquiteturas voltadas a método de detecção de *outliers* baseado em estatística e a exploração espaço-temporal para localização dos nodos para a detecção do evento. Para evidenciar as contribuições da proposta, apresenta-se na Tabela 2 os requisitos utilizados para classificação baseados nos trabalhos de (RASSAM; ZAINAL; MAAROF, 2013b), (ZHANG; MERATNIA; HAVINGA, 2010) e (BHOJANNAWAR; BULLA; DANAWADE, 2013).

Os trabalhos que envolvem apenas detecção de *outliers* não necessitam de correlação espaço-temporal. Por outro lado, a detecção de eventos utiliza a correlação espaço-temporal na sua identificação. Os trabalhos comentados a seguir são os que possuem maior proximidade com a proposta deste estudo.

Em relação ao trabalho de Pei et al. (2014), a proposta diferencia-se no método proposto para determinar a proximidade

Tabela 2 – Comparações entre arquiteturas.

Trabalhos	Entrada de dados	Modelo de Estrutura	Modelo de Operação	Adaptado a Mudanças	Correlação	Agrupamento dos dados	Formação da Rede
Sharma, 2010a	Univariado	Distribuído	Offline	Não	Não	Não	Não
Li, 2010	Multivariado	Centralizado	Offline	Não	Não	Não	Não
Xei, 2012	Univariado	Distribuído	Online	Não	Não	Não	Não
Yao, 2010	Univariado	Distribuído	Offline	Não	Não	Não	Não
Aggarwal, 2001	Multivariado	Centralizado	Offline	Não	Espaço-temporal	Não	Não
Oladimeji, 2015	Univariado	Centralizado	Online	Sim	Não	Sim	Não
Bettencourt, 2007	Multivariado	Distribuído	Online	Sim	Espaço-temporal	Não	Não
Pei, 2014	Univariado	Distribuído	Online	Sim	Espaço-temporal	Não	Não
Mousavi, 2013	Univariado	Distribuído	Online	Sim	Espaço-temporal	Não	Não
Proposta	**Univariado**	**Distribuído**	**Online**	**Sim**	**Espaço-temporal**	**Sim**	**Sim**

dos nodos, embora em comum utilize a potência de sinal como variável. Esta será medida e relacionada com os vizinhos gerando uma tabela de proximidade. A correlação temporal proposta não necessita de sincronismo entre os nodos.

Em relação ao trabalho de Mousavi et al. (2013), a proposta diferencia-se na formação do grafo. Ela utiliza a potência do sinal para localização, enquanto o trabalho de Mousavi necessita ter o conhecimento exato da posição dos nodos no cenário. Em relação a correlação temporal, a proposta não faz uso de cadeia de Markov para modelar o comportamento das leituras.

Considerando as diferenças da proposta deste trabalho com o trabalho de Oladimeji, Smiee e Mieee (2015), os métodos utilizados para identificação do evento e do tratamento das anomalias são distintos. Com relação à identificação de eventos, Oladimeji, Smiee e Mieee (2015) aplicam a técnica *k-means* para detecção de eventos; a qual consiste em agrupar os dados por similaridades desconsiderando a localização dos nodos no cenário. Por outro lado, na proposta desta tese considera-se a correlação espacial dos nodos para determinação do evento, visando melhor precisão na identificação dos eventos. Quanto ao método de tratamento das anomalias, Oladimeji, Smiee e Mieee (2015) utilizam redes neurais (abordagem supervisionada) com um classificador previamente disponível. Nesta proposta optou-ser pelo método baseado em estatística sem disponibilidade de informações prévias (abordagem não supervisionada). Esta escolha baseia-se no reduzido consumo dos recursos e na diminuição da transmissão dos dados.

Com relação ao trabalho de Bettencourt (2007), a

proposta tem o diferencial de não necessitar o conhecimento das leituras anteriores e também não fazer teste de significância sobre a distribuição dos dados. Desta forma, a proposta com relação ao requisito temporal é mais tolerante à criticidade das aplicações.

3.6.2 Trabalhos Correlatos - Formação de Cluster-Tree para RSSF

Quando RSSF são implementadas em cenários de larga escala, não é realístico assumir que a comunicação de cada sensor alcance seu destino em apenas um salto, sendo necessário adotar a abordagem de comunicação *multi-hop*, tal como topologia hierárquica (FELSKE et al., 2013). Neste contexto, topologia *cluster-tree* e o escalonamento do beacon são as soluções mais atrativas (LEAO et al., 2016; LEÃO et al., 2017). Essas técnicas visam melhorar o desempenho da comunicação em rede, permitindo aos grupos de nodos estabelecerem seus próprios ciclos de trabalho, constituídos de períodos ativos e ociosos, reduzindo colisões de quadros, período de escuta inativo e problemas de escuta.

Geralmente, uma topologia *cluster-tree* é construída pelo agrupamento de nodos vizinhos de acordo com critérios específicos. Um exemplo é o Algoritmo de Coleta de Dados Baseado em Cluster-Tree (TCBDGA) (ZHU et al., 2015) cuja a ideia básica é construir uma rede *cluster-tree* considerando características específicas dos nodos, tais como sua energia residual, sua distância à estação base e o número de vizinhos. Posteriormente, a rede é decomposta em várias subárvores e um coletor móvel que é responsável por coletar dados dos

nodos.

Outros exemplos são os trabalhos (DING; TIAN; YU, 2016) e (HONG; WANG; LI, 2016), que consideram o valor do RSSI e a energia residual dos nodos para formar os *clusters*. Em Bholowalia e Kumar (2014), além do RSSI e da energia residual dos nodos, o posicionamento geográfico dos nodos também é usado para formar os *clusters*. Tanto nesse algoritmo quanto no TCBDGA, a abordagem de reconhecimento de localização implica em maior consumo de energia e complexidade.

Outros trabalhos formam *cluster-tree* levando em consideração o roteamento das mensagens para a estação base. Por exemplo, uma abordagem baseada em cálculos geométricos é proposta em (XIE; JIA, 2014), em que os nodos mais próximos dos centros das áreas mais "povoadas" se tornam *cluster-head* (CH), e o resto dos nodos está associado ao CH que tem o menor número de saltos para alcançar a estação base.

Em Khatiri, Mirjalily e Khademzadeh (2012), cada nodo mantém uma tabela de vizinhança com informações relevantes sobre os nodos vizinhos como: profundidade na árvore, qualidade do enlace e tipo do dispositivo. O algoritmo proposto define o caminho mais curto usando três critérios: contagem mínima de saltos, congestionamento mínimo e qualidade máxima do enlace.

Em Kim, Bang e Lee (2014), um algoritmo inspeciona sua tabela de vizinhança para encontrar um nodo com um caminho de árvore mais curto para o nodo de destino e o seleciona como próximo salto. Embora esse algoritmo defina o caminho o mais curto para a estação base, ele não considera

nenhuma informação sobre o escalonamento da rede nem sobre como a comunicação de dados é realizada.

Nesta tese, argumenta-se que, quando o monitoramento é feito para detectar eventos relevantes em uma área ampla, ignorar os valores de dados detectados durante o estabelecimento da topologia de rede é certamente uma abordagem ineficaz. Em RSSF de larga escala, o uso de técnicas analíticas de "big data", como fusão de informação (PINTO et al., 2014), clusterização de dados (JAIN, 2010) e tratamento de *outliers* (ANDRADE et al., 2016), é fundamental quando se manipula com grandes quantidades de dados detectados.

Nosso argumento é que a topologia de comunicação de rede deve ser integrada a essas técnicas. Frequentemente, os dados dos nodos podem ser separados em grupos distintos, como dados de nodos normais ou com falhas (HE; WANG, 2007), ou dados de um grupo de nodos monitorando a temperatura na sombra ou em áreas ensolaradas (BOANO et al., 2010) ou de um grupo de nodos que monitoram alta ou baixa concentração de gás tóxico (SHU; MUKHERJEE; WU, 2016) etc. O agrupamento ou clusterização de dados é exatamente o estudo de métodos e algoritmos para agrupar dados de acordo com sua similaridade e, portanto, as técnicas de agrupamento de dados também devem ser usadas para orientar a formação da topologia da rede.

Cada um dos grupos criados por técnicas de agrupamento de dados (por exemplo, *k-means*) (JAIN, 2010), aproximação de k-vizinhos mais próximos (HE; WANG, 2007) ou (CHEN; ALMEIDA; WANG, 2010) frequentemente exigem diferentes periodicidades na coleta de seus dados. Por exemplo,

os *clusters* que monitoram dados relacionados a nodos de operação defeituosos ou dados com concentração de temperatura/gás acima do normal normalmente exigirão uma frequência de monitoramento muito maior do que os dados de monitoramento de cluster no estado estacionário. Além disso, o processamento na rede (por exemplo, técnicas de fusão de dados) também é comumente aplicado nessas redes, nas quais os dados são processados nos próprios nodos ao longo do caminho antes de chegar à estação base.

Nesse sentido, seria aconselhável ter uma topologia de comunicação que separasse explicitamente os caminhos de comunicação que possuem nodos com diferentes grupos de dados. Como esses grupos lógicos são geralmente compostos por nodos geograficamente distantes, isso motiva abordagens de formação de topologia que considerem a semelhança entre os dados detectados no processo de criação dos caminhos de comunicação.

A ideia subjacente deste trabalho é demonstrar que uma abordagem básica (baseline) para realizar formação topológica (por exemplo, baseada em uma estratégia de topologia) *cluster-tree* ao comparar com uma organização de agrupamentos de dados lógicos (por exemplo, formados por técnica *k-means*) é ineficiente em termos de consumo de energia e/ou tempo de resposta na detecção de eventos. Uma abordagem que integra os conceitos de *clusters* hierárquicos em RSSF de grande escala, juntamente com a análise de agrupamentos de dados, é necessária.

Apesar da forte relação entre a pesquisa de detecção de *big data* e a formação de *cluster-tree* em RSSF de larga escala,

a fusão das duas áreas de pesquisa em uma única abordagem não é automática. Apenas unir nodos geograficamente distantes no mesmo *cluster* pode ser inútil do ponto vista de roteamento da comunicação RSSF. Por outro lado, também não é eficaz juntar nodos próximos no mesmo *cluster*, se eles tiverem dados com características diferentes, exigindo, por exemplo, que sejam amostrados em intervalos diferentes.

A Figura 14, apresenta a taxonomia do trabalho de (LI et al., 2011) para o roteamento em RSSF de larga escala. Nela foi inserida uma nova categoria baseada nos dados para formação de topologia em RSSF de larga escala, objeto desta tese.

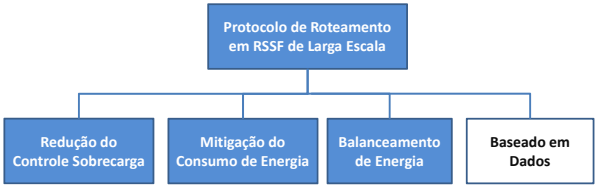


Figura 14 – Taxonomia para protocolos de roteamento em RSSF de larga escala. Adaptado de: (LI et al., 2011).

Uma das contribuições desta tese é propor uma abordagem inovadora para formação de redes *cluster-tree*, chamada de DbCTF (Formação de Redes *cluster-tree* baseada em Dados).

3.7 Considerações do Capítulo

Este capítulo descreveu os conceitos fundamentais necessários para o entendimento da detecção de *outliers* em RSSF, e também as definições sobre detecção de eventos essenciais ao processo de identificação do evento relevante. Os

requisitos para classificação dos métodos foram apresentados, onde seus parâmetros são aplicados para comparar características de diferentes técnicas de detecção. Ademais, tratou-se dos métodos de detecção imprescindíveis em termos de eficiência do recursos para o cenário aplicado.

Também neste capítulo foram apresentados alguns dos trabalhos que influenciaram esta proposta de tese. Os trabalhos descritos focam em abordagens estatísticas por serem mais leves e tratáveis em RSSF. Além disso, buscou-se comparar a proposta com trabalhos que apresentassem abordagens que tratam a relação espaço-temporal entre os dados analisados.

Por fim, foi discutido como a abordagem de formação de redes *cluster-tree* baseada em dados se insere na literatura e sua comparação com os trabalhos relacionados.

4 Arquitetura para Detecção, Identificação e Tratamento de *Outliers* em RSSF de larga Escala

4.1 Introdução

Conforme descrito no Capítulo 3, dados anômalos ou *outliers* são aqueles cujos valores não estão em conformidade com o "padrão normal" de comportamento de outros dados detectados (CHANDOLA; BANERJEE; KUMAR, 2009). Aplicações que utilizam RSSF de larga escala muitas vezes operam em ambientes hostis e sem informações prévias, sendo sujeitas a falhas, ataques maliciosos, além de leituras imprecisas e não confiáveis dos seus sensores. Portanto, a existência desses tipos de dados é inerente a essas aplicações. Por outro lado, um *outlier* detectado pode ser resultante de algum evento relevante que começou a ocorrer na área monitorada e que ainda não foi detectado pela maioria dos nodos. Por conseguinte, além da detecção de *outliers*, a identificação/classificação destes é uma etapa essencial para seus adequados tratamentos.

No contexto das RSSF, as técnicas de detecção, identificação e tratamento de *outliers* para RSSF precisam observar suas especificidades, tais como, recursos

computacionais limitados, consumo energético, natureza dinâmica de sua topologia, alcance limitado das transmissões e perdas de mensagens na comunicação. Para exacerbar o problema de lidar com recursos limitados, as RSSF precisam operar por longos períodos sem a necessidade de intervenção humana. Portanto, existe a necessidade dessas redes possuírem capacidades de autogerenciamento, autoreconfiguração, auto-otimização e autocura (KEPHART; CHESS, 2003).

A partir dessas características, torna-se claro que muitas técnicas tradicionais para detecção, identificação e tratamento de *outliers* não são diretamente aplicáveis no contexto das RSSF. Além disso, no sentido de promover uma maior eficiência no tratamento dos *outliers* e detecção de eventos no contexto de RSSF de larga escala, essas técnicas não deveriam atuar de forma isolada. Contudo, conforme visto no capítulo anterior, os trabalhos da literatura relacionados a esta tese deixam uma lacuna na integração dos métodos de agrupamento de dados e detecção de *outliers*, além de desconsiderarem a formação topológica da rede. A integração dessas técnicas, em um contexto de RSSF de larga escala, onde a topologia precisa mudar ao longo do tempo para acompanhar as mudanças nos dados monitorados, é a motivação desta proposta.

Nesta tese é proposta uma arquitetura direcionada a detectar e identificar adequadamente *outliers* em RSSF de larga escala. Os principais objetivos da arquitetura são os de garantir (i) a qualidade dos dados monitorados e entregues pela aplicação, (ii) uma correta fusão dos dados monitorados, e (iii) uma detecção confiável e responsiva de eventos relevantes que

ocorram na rede. Ademais, considerando a natureza das RSSF e dos dados monitorados, a arquitetura (iv) busca otimizar o uso dos recursos da rede e dos nodos, além de (v) possuir característica autonômica, permitindo o acoplamento de uma estratégia para o autoajuste da topologia da rede, conforme a mudança dinâmica dos dados monitorados.

4.2 Justificativas da Arquitetura

O monitoramento de grandes áreas pode envolver vários tipos de cenários, tais como agricultura de precisão, pecuária, indústria e monitoramento ambiental (AIKENETA, 2002). Em diversas dessas aplicações de monitoramento, a operação é realizada de forma centralizada, tendo um dispositivo sensor específico para cada grandeza. Nestes casos, por terem apenas um sensor, este deverá ser preciso, confiável, durável e, em razão destes fatores, geralmente dispendioso. Nessa abordagem centralizada podem ser destacados problemas relacionados à confiabilidade da aplicação ser baseada num único sensor, e também no fato do armazenamento dos dados geralmente ser feito localmente. Basear a confiabilidade dos dados em medições de um único sensor acarreta em baixa tolerância a faltas, principalmente porque essa confiabilidade fica sujeita a problemas de instalação deste dispositivo sensor.

Sobre o armazenamento local, geralmente feito em dispositivos *data-logger*, este envolve custos de manutenção para retirada e substituição das mídias dos dados, uma vez que o local pode ser de difícil acesso. Além disso, esse tipo de armazenamento é adequado apenas para abordagens que atuam *offline* (ex. análise *a posteriori* dos dados), dificultado

qualquer abordagem que atue de forma *online* (ex. monitoramento com geração de alarmes).

As RSSF surgem com uma alternativa a soluções centralizadas. Nessas redes, vários sensores de baixo custo monitoram as grandezas do ambiente de forma cooperativa através de nodos sensores de baixo custo, e encaminham seus dados para uma estação base. Abordagens *on-line* e dinâmicas podem ser empregadas, podendo haver não só monitoramento, mas também atuação em tempo real sobre o ambiente, caso alguns nodos da rede possuam essa capacidade. Além disso, o próprio monitoramento pode atuar de forma adaptativa, ajustando dinamicamente a periodicidade das amostragens e envios de mensagens em regiões específicas da rede, focando aquelas onde eventos relevantes estejam prestes a ocorrer.

A confiabilidade dos dados obtida nas RSSF emprega técnicas adequadas para a detecção e remoção de *outliers* oriundos de dados espúrios (ALCARAZ et al., 2010) (AKYILDIZ et al., 2002). Neste trabalho, através de uma análise sistemática da literatura (Apêndice A), foram observadas as proposições de muitas técnicas de detecção de *outliers* que assumem que estes são resultantes de sensores faltosos ou erros na transmissão de dados. Esses trabalhos, portanto, não possuem uma etapa de identificação dos *outliers* pois assumem como premissa que estes são dados espúrios gerados na rede e, geralmente, focam em técnicas para a sua remoção.

Algumas poucas pesquisas focam na identificação e tratamento de eventos. No sentido de identificar o evento, estas geralmente assumem o conhecimento prévio da posição de cada

nodo, buscando uma correlação espacial nos *outliers* detectados. Por outro lado, o conhecimento da localização dos nodos em RSSF possui um alto custo associado, seja relacionado ao custo da implantação de nodos cuidadosamente em pontos específicos na área monitorada, seja pelo custo da adição de dispositivos GPS em cada nodo sensor. Por este motivo, a arquitetura proposta no presente trabalho tem como premissa o desconhecimento na localização dos nodos. Contudo, esse pressuposto exacerba o problema da identificação dos *outliers* pois, na proposta desta tese, essa etapa passa a ser efetuada apenas através da Potência do Sinal Recebido (RSSI) gerada durante a comunicação entre os nodos.

A arquitetura busca configurar a topologia conforme os próprios dados monitorados no ambiente, de forma a tornar adequada a periodicidade e os atrasos fim-a-fim das mensagens aos prazos (*deadlines*) que estejam eventualmente associados ao tipo de dado monitorado. Um dos desafios enfrentados pela arquitetura é que os valores lidos pelos nodos podem mudar ao longo do tempo. Isso torna inadequado o uso de topologias de comunicação estáticas, já que estas não estarão preparadas para o atendimento de prazos e periodicidades associados aos eventos.

Resumidamente, um dos objetivos da arquitetura proposta é que esta seja independente com relação a: localização prévia dos nodos; densidade dos nodos na área (balanceada, densa ou esparsa); além da distribuição dos nodos na área monitorada (regular ou irregular).

Na sequência, na Seção 4.3 são descritos os principais pressupostos da arquitetura. Em seguida, a arquitetura é

detalhada na Seção 4.4, sendo descritos os métodos e funcionalidades que a compõe.

4.3 Pressupostos da Arquitetura

Nesse trabalho assumimos um sistema formado por um conjunto de nodos que serão organizados na forma de uma árvore de *clusters* com objetivo de monitorar uma área. O sistema possui subjacente um modelo de rede e de aplicação, e os seus principais pressupostos são enumerados a seguir:

p1 – Há um grande número de nodos sensores que são implantados na área a ser monitorada, são estáticos e não são cientes de suas posições;

p2 – Os dados enviados pelos nodos geralmente não conseguem alcançar o nodo coordenador com apenas um salto;

p3 – Os nodos sensores são homogêneos (com os mesmos tipos de sensores, rádio, capacidade de processamento, bateria e memória);

p4 – Os canais de comunicação são simétricos com relação à intensidade de sinal;

p5 – A aplicação é *convergecast* e periódica, cuja periodicidade é dependente do tipo ou do valor do dado monitorado;

p6 – Os conjuntos de dados monitorados formam agrupamentos que podem ser identificados através de uma técnica de clusterização de dados adequada;

p7 – Os agrupamentos de dados formados podem eventualmente mudar ao longo do tempo;

p8 – Os dados anômalos lidos pelos nodos sensores podem ser identificados, através de técnicas adequadas, como espúrios ou resultantes de algum evento relevante na rede;

p9 – Os dados enviados pelos nodos sensores serão tratados no coordenador e podem ter prazos para alcançá-lo.

O pressuposto **p1** representa especificamente a condição dos nodos geralmente serem implantados de forma aleatória, devido a dificuldades e custos de implantação individual de cada nodo, além do custo do hardware necessário para determinação da posição (ex. receptores GPS). O pressuposto **p2** é usual em aplicações de monitoramento de grandes áreas usando RSSFs. Apesar da especificação do padrão IEEE 802.15.4 permitir nodos com capacidades diferentes, através da distinção entre dispositivos RFD e FFD, o pressuposto **p3** permite o uso de sensores de baixo custo e maior flexibilidade no método de formação de topologia, possibilitando que qualquer nodo assuma o papel de CH.

O pressuposto **p4** é justificado pelo uso do valor de RSSI, o qual pode ser medido por qualquer nodo no momento da recepção de um *frame*. O RSSI é uma métrica adequada para avaliar a qualidade do canal de comunicação, e seu valor pode ser assumido como simétrico nos rádios atualmente usados (JIN et al., 2015) (YAO et al., 2015). O pressuposto **p5** é usual em aplicações de monitoramento em RSSF com grandes dimensões geográficas, onde nodos sensores coletam dados periodicamente e os enviam ao coordenador PAN através de um comportamento *convergecast*. Neste tipo de aplicação eventualmente ocorrem transmissões de mensagens pelo coordenador destinadas para todos os nodos da rede

(*broadcast*), tais como mensagens de reconfiguração da rede. Contudo, assume-se que esse tipo de mensagem é mais raro de ocorrer na rede e, por esse motivo, não é otimizado na arquitetura proposta.

Os pressupostos **p6**, **p7**, **p8** e **p9** são relacionados aos dados monitorados na rede pelos nodos sensores. Assume-se que esses dados possuem correlações entre si, as quais são mantidas durante um período de tempo. Essas correlações podem ser usadas em técnicas para a identificação e formação de agrupamentos de dados (*data clustering*) tais como o *k-means* e suas variações, as quais são amplamente usadas em trabalhos para análise e tratamento de grande massa de dados (*big data*). Eventualmente, dados anômalos são detectados entre os dados monitorados e devidamente identificados. Essa identificação poderia ser feita no interior da própria rede (*in-network processing*), por exemplo, em nodos CHs. Contudo, a arquitetura proposta neste trabalho assume que o coordenador da rede irá receber os dados enviados através de comunicação *convergecast* e, de forma centralizada, formar os agrupamentos de dados e identificar os *outliers*. Para o caso do *outlier* ser resultante de um evento que começa a ser detectado na rede, importante que este seja identificado o quanto antes. Portanto, assume-se no modelo adotado que os dados podem possuir prazos máximos (*deadlines*) associados à entrega na estação base.

4.4 Descrição da Arquitetura

Esta seção descreve a arquitetura proposta para detecção, identificação e tratamento de *outliers* em RSSF de

larga escala, apresentada na Figura 15. A integração dos métodos aplicados nesta arquitetura confere vantagens quando comparada aos mesmos individualmente. Considerando os impactos da topologia de comunicação sobre os resultados obtidos no próprio monitoramento da rede e, consequentemente, na formação dos grupos de dados e na detecção de *outliers*. Esta arquitetura, de forma inovadora, agrega uma estratégia de reconfiguração dinâmica da topologia *cluster-tree* da rede baseada nos próprios dados monitorados.

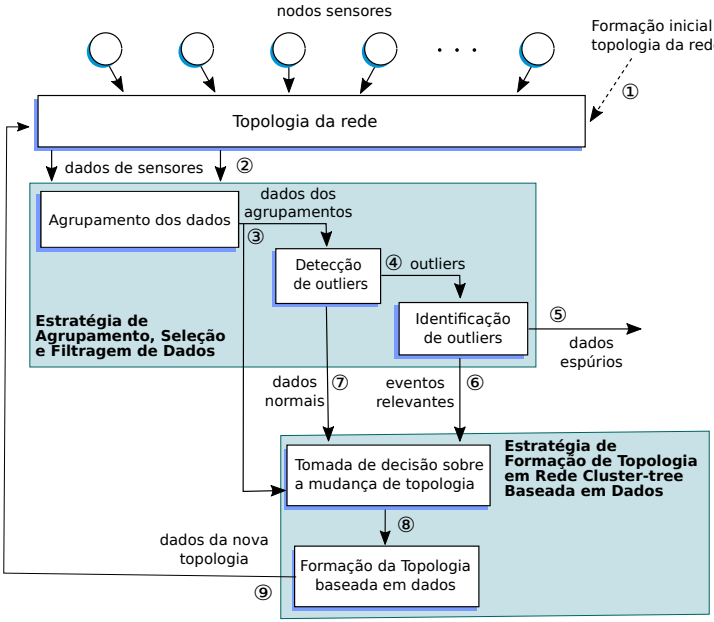


Figura 15 – Arquitetura para detecção, identificação e tratamento de eventos em RSSF de larga escala.

A escolha dos métodos usados na arquitetura considera os seus custos-benefícios, sendo os principais métodos

da arquitetura aqueles responsáveis pela: (i) clusterização dos dados; (ii) detecção do *outliers*; (iii) detecção de eventos; e (iv) formação dinâmica de topologia *cluster-tree* baseada em dados para redes de larga escala.

Para efeitos de comparação com outras abordagens, pretende-se analisar: a eficiência da arquitetura quanto à detecção de eventos e quanto à estratégia de formação das rotas para redes *cluster-tree*; as métricas de desempenho da rede com respeito a atrasos fim-a-fim, taxas de sucesso e consumos energéticos dos nodos; e a adaptação às mudanças dinâmicas da rede e do ambiente monitorado, sem conhecimento de posições dos nodos e sem informações prévias do ambiente.

A arquitetura ilustrada na Figura 15 é executados na estação base e pode ser também executados nos *Cluster-head*. Na sequencia são descritos os componentes da arquitetura, sendo divididos em topologia da rede; agrupamento, seleção e filtragem dos dados; detecção de *outliers*; identificação de *outliers*; tomada de decisão; e formação de rede baseada em dados.

4.4.1 Topologia da Rede

A topologia da rede – assim como os nodos que a compõe – não é propriamente um componente da arquitetura, mas sim um atributo do ambiente com o qual a arquitetura interage. Como a arquitetura foi criada de forma a interagir com o ambiente e reconfigurar a topologia de comunicação de forma autonômica, essa topologia é representada graficamente na Figura 15 na forma abstrata de um retângulo, de forma a facilitar a compreensão do funcionamento da arquitetura.

Assume-se que, inicialmente, há uma formação

topológica da rede, representada pelo evento ① na Figura 15. Essa formação inicial provavelmente não é otimizada pois é anterior ao funcionamento da arquitetura. Assume-se que essa formação inicial *ad-hoc* seja suficiente para que os dados dos sensores alcancem a estação base, onde as estratégias propostas para a arquitetura irão executar.

Os nodos da rede trocam informações entre si, a fim de descobrir seus vizinhos, pois a definição da vizinhança é uma informação fundamental para a posterior redefinição topológica da rede. Portanto, o evento ② da Figura 15 representa não somente o envio periódico de dados coletados pelos nodos, mas também o envio da matriz de conectividade, a qual descreve a vizinhança de cada nodo.

4.4.2 Estratégia de Agrupamento, Seleção e Filtragem dos Dados

Esse módulo da arquitetura é formado pelos seguintes componentes: (i) Agrupamento dos Dados, (ii) Detecção de *Outliers*, e (iii) Identificação de *Outliers*.

Agrupamento dos Dados

O método selecionado para a classificação dos dados da arquitetura analisou o seu custo-benefício computacional e sua adaptabilidade ao cenário. Os métodos mais viáveis para lidar com conjuntos de dados no contexto de RSSF aplicam técnicas de probabilidade, estatística ou clusterização (ZHANG; MERATNIA; HAVINGA, 2010).

O processo de clusterização utiliza dados dos nodos, os

quais são classificados por semelhanças desconsiderando suas localizações e sem informações prévias. Na arquitetura proposta foi selecionado o método denominado *k-means*. Segundo Žalik (2008), o *k-means* é um algoritmo de mineração de dados simples, sem supervisão que analisa e classifica dados numéricos automaticamente. Uma desvantagem conhecida dessa técnica, é que o número de grupos precisa ser informado previamente à formação dos agrupamentos. A utilização da abordagem de clusterização permite formar agrupamentos segundo a semelhança dos dados antes da retirada de *outliers* e do cálculo da média. Deste modo, essa abordagem confere maior precisão e confiabilidade à amostra. A Seção 5.3, apresenta um estudo sobre a melhoria da precisão nos dados utilizando a técnica de clusterização *k-means* e o método de detecção de *outliers* baseado em estatística.

Ao fim do processo de agrupamento dos dados, os grupos são gerados e os dados agrupados (e.g. sua média) são encaminhados para a etapa detecção de *outliers* (assinalado como ③ na Figura 15). Paralelamente, a informação sobre os agrupamentos gerados também é enviada para o componente de tomada de decisão para que este decida sobre a necessidade da mudança na topologia da rede.

Detecção de *Outliers*

Esse módulo tem como objetivo detectar anomalias em um conjunto de dados, nos quais todas as instâncias recebem o mesmo tratamento independente da sua região. Dessa forma, é necessário observar quais regiões com diferentes características e densidades de dados podem gerar anomalias.

Quatro requisitos foram importantes para definir a escolha do método de detecção de *outliers* baseado em estatística (ZHANG; MERATNIA; HAVINGA, 2010), (RASSAM; ZAINAL; MAAROF, 2013a), (BHOJANNAWAR; BULLA; DANAWADE, 2013): baixo custo computacional; adequação ao cenário dinâmico e distribuído das RSSF; capacidade de operar sem disponibilidade prévia dos dados; e atendimento de anomalias com requisitos temporais.

Em suma, para a seleção do método detecção estatístico foram considerados os aspectos de eficiência na precisão e na taxa de detecção de *outliers* e eventos relevantes. Embora importantes, neste trabalho não foram avaliados outros aspectos de eficiência dos métodos como consumo de energia, utilização de memória e processador. Com base em estudos da literatura (ZHANG; MERATNIA; HAVINGA, 2010), (RASSAM; ZAINAL; MAAROF, 2013a), (BHOJANNAWAR; BULLA; DANAWADE, 2013), optou-se pela escolha dos métodos estatísticos.

Sobre a abordagem baseada em estatística, alguns métodos de detecção de *outliers* foram avaliados e serão mostrados na Seção 5.3 como: critério de Chauvenet (TAYLOR, 2012), critério de Pierce (ROSS; PH, 2003), método de Marzullo (MARZULLO, 1990) e método de Elmenreich (ELMENREICH, 2002).

Como resultado, o módulo de detecção de *outliers* entrega duas possibilidades:

- *Outlier* detectado, sendo necessária uma ação de identificação de sua natureza para posterior tratamento.

Esse evento é representado por ④ na Figura 15.

- Dado normal, livre de anomalias, podendo ser utilizado na tomada de decisão para mudança de topologia. Esse evento representado por ⑦ na mesma figura.

Ao fim deste processo, com a correta detecção de *outliers*, o cálculo das médias dos valores dos sensores usando apenas dados normais possibilita uma melhoria nas tomadas de decisão feitas na estação base. A forma de utilização dessas médias depende da aplicação que usa a arquitetura proposta e, por isso, ela não é explicitamente representada na Figura 15.

Identificação de *Outliers*

Esse processo visa a identificação de eventos para seu posterior tratamento. No processo de identificação é utilizada a correlação espaço-temporal dos nodos no cenário para diferenciar os dados espúrios dos eventos relevantes, essa diferenciação é representada na Figura 15 pelos eventos ⑤ e ⑥, respectivamente.

A detecção de eventos parte, inicialmente, da detecção de um *outlier* individual em um nodo e, na sequência, compara sua leitura com a dos seus vizinhos em um intervalo espaço-temporal determinado. Quando detectada uma anomalia, a localização do nodo deve ser considerada e relacionada com seus vizinhos, além do instante de tempo da detecção. Contudo, um aspecto relevante é a questão da dificuldade de se efetuar essa correlação temporal e espacial, pois a arquitetura inicia sem informações prévias da implantação dos nodos no cenário, usando apenas a matriz de

vizinhança como critério de identificação do alcance da comunicação de cada nodo.

Após a passagem pelo processo de identificação, quando um *outlier* é classificado como um evento relevante, essa informação é repassada para módulo de tomada de decisão sobre a mudança de topologia da arquitetura (evento ⑥ na Figura 15). Importante destacar que a aplicação que utiliza essa arquitetura irá efetuar um tratamento relacionado à detecção desse evento como, por exemplo, a geração de alarmes. Contudo, como essa ação é dependente da aplicação de monitoramento, ela não é representada explicitamente na arquitetura proposta ilustrada na Figura 15.

4.4.3 Estratégia de Formação para Redes *Cluster-Tree* Baseada em Dados

Esse módulo da arquitetura é formado por dois componentes: (i) Tomada de Decisão para Mudança de Topologia e (ii) Formação de Topologia Baseada em Dados.

Tomada de Decisão para Mudança de Topologia

Após a etapa de detecção de *outliers*, dados classificados como resultantes de um evento relevante, além daqueles considerados como "normais", são encaminhados ao processo de formação de topologia baseada em dados (eventos representados por ⑥ e ⑦ na Figura 15). Nessa etapa, esses dados são analisados para uma tomada de decisão por parte desse módulo.

A base para a decisão sobre a mudança da topologia é

configurável na arquitetura. Essa mudança poderia, por exemplo, ser periódica e ocorrer através do disparo de um temporizador, sem a necessidade das informações representadas na Figura 15 por ⑥ e ⑦. Outra alternativa, seria disparar uma mudança na topologia a cada vez que houvesse qualquer mudança nas informações dos agrupamentos de dados (informação representada por ③).

Contudo, disparar a mudança de topologia de forma periódica, não é uma solução eficiente, pois se não houver mudanças nos agrupamentos de dados formados, não há necessidade de se alterar a topologia. Além disso, disparar a mudança de topologia a cada mudança na informação de agrupamento, também pode ser ineficiente, pois a qualquer mínima mudança (por exemplo, apenas um nodo mudou de agrupamento) irá disparar um processo com muitas etapas e trocas de mensagens entre os nodos.

Por esses motivos, no Capítulo 6 desta tese de doutorado, o uso de outra alternativa foi avaliado através de experimentos simulados. Em vez de disparar uma mudança de topologia a cada mudança na informação dos agrupamentos, nos experimentos foi utilizada a informação de um percentual de variação dos nodos nos grupos gerados pela técnica *k-means*. Além disso, cada mudança que ocorrer na topologia passa a ser válida por um período de tempo. Ambos os valores do percentual e do período de tempo, usados para essa tomada de decisão sobre a topologia, são parâmetros pré-estabelecidos e configuráveis.

Resumidamente, a alternativa avaliada nesta tese possui o seguinte funcionamento: quando uma nova informação

é recebida por este módulo (evento ③) e quando houver passado um determinado período de tempo desde a última mudança na topologia (ex. 20 minutos), os nodos pertencentes aos novos conjuntos dos agrupamentos são comparados com os dos agrupamentos formados anteriormente. Quando o percentual de mudança ficar acima de um limite pré-estabelecido (e.g 20% dos nodos mudaram de agrupamento), os dados recebidos são enviados para o módulo de formação da topologia baseada em dados (evento ⑧ na Figura 15) para que este módulo calcule a nova formação topológica. Nesta tese, a formação da nova topologia é efetuada pela heurística denominada DbCTF, a qual é descrita em pormenores na Seção 6.3.2.

Formação de Topologia Baseada em Dados

No processo de formação de topologia, os nodos trabalham para construir rotas baseadas nos dados. Como o modelo de comunicação assume que a comunicação é *convergecast*, o objetivo é formar caminhos desde os sensores até a estação base, separando as rotas conforme os agrupamentos de dados. Desta forma, é possível, por exemplo, se estabelecer parâmetros na rede distintos conforme os diferentes agrupamentos.

As rotas criadas pela mudança de topologia da rede *cluster-tree* podem ter diferentes taxas de monitoramento em virtude de haver, por exemplo, um evento relevante monitorado em um dos *clusters* da árvore pertencentes àquela rota. Consequentemente, os atrasos fim-a-fim, consumos energéticos e taxas de transmissão serão influenciados conforme a taxa de

monitoramento daquele grupo específico.

4.5 Métricas de Desempenho

Apesar de haver muitas arquiteturas voltadas para detecção de *outliers* e/ou fusão de dados, os módulos responsáveis pela estratégia de formação de topologia em redes *cluster-tree* baseada em dados conferem uma natureza inovadora na arquitetura proposta, pois a revisão sistemática na literatura, feita no contexto deste trabalho, não encontrou proposta similar que reconfigura a topologia da rede baseada nos próprios dados monitorados. Essa natureza inovadora reforça a necessidade de se fazer uma avaliação de desempenho da arquitetura e de seus módulos escolhidos para implementarem suas estratégias.

Para escolha dos parâmetros de avaliação levou-se em consideração os essenciais ao contexto desta pesquisa e recorrentes na literatura. Para avaliação da estratégia de agrupamento, seleção e filtragem dos dados, as métricas escolhidas foram:

- Detecção de *outliers*: número de dados anômalos detectados;
- Detecção de eventos: número de eventos relevantes identificados;
- Dados Espúrios: número de dados descartados;

As métricas selecionadas para avaliação da estratégia de formação para redes *cluster-tree* baseada em dados foram:

- Atraso fim a fim: tempo decorrido desde a geração da mensagem no nodo até a chegada na estação base;
- Tempo de vida do nodo: tempo estimados da bateria do nodo;
- Taxa de sucesso na entrega de pacotes: percentual médio de mensagem entregues com sucesso na estação base;

Tais métricas possibilitam avaliar as estratégias propostas e seus resultados são discutidos nos Capítulos 5 e 6.

4.6 Considerações do Capítulo

Neste capítulo foram detalhados os componentes da arquitetura proposta, cuja concepção modular busca alcançar uma melhor escalabilidade, flexibilidade, precisão na fusão de dados e eficiência na detecção de eventos. Os componentes da *Estratégia de Agrupamento, Seleção e Filtragem de Dados* contribuem para melhorar a precisão na detecção de *outliers* e na detecção de evento relevantes. Complementarmente, os componentes da *Estratégia de Formação de Topologia em Rede Cluster-tree Baseada em Dados* promovem o melhor atendimento aos requisitos do evento no atraso fim-a-fim, no consumo energético da rede e na taxa de sucesso dos pacotes.

Os próximos capítulos descrevem os experimentos efetuados para avaliar a arquitetura proposta, através de dois cenários distintos. No primeiro, o componente *Estratégia de Agrupamento, Seleção e Filtragem de Dados* da arquitetura é avaliada em um cenário usando nodos reais, ilustrando o comportamento da arquitetura em uma aplicação de

monitoramento de temperatura e umidade do ar. Nesse cenário, descrito no Capítulo 5, não houve preocupação com a escalabilidade e mudanças de topologia da rede. O enfoque foi na verificação de técnicas que exigem pouca capacidade computacional da rede e que, portanto, podem ser empregadas em RSSF.

No segundo cenário, o componente *Estratégia de Formação de Topologia em Rede Cluster-tree Baseada em Dados* da arquitetura é avaliado e seus resultados descritos no Capítulo 6. Dessa vez, a avaliação é feita através de simulação, o que permitiu a geração de diversos cenários com até uma centena de nodos. O enfoque foi na avaliação da adequação da arquitetura em cenários com grandes dimensões, maior número de nodos, e com topologia da rede que precisa se ajustar dinamicamente para se adequar a mudanças nos valores dos dados monitorados.

5 Estratégia de Agrupamento, Seleção e Filtragem dos Dados

5.1 Introdução

Este capítulo avalia os componentes da arquitetura responsáveis pela estratégia de agrupamento, seleção e filtragem de dados destacados na Figura 16. No sentido de efetuar a avaliação, foram efetuados diversos experimentos envolvendo uma aplicação de RSSF voltada para o monitoramento do ambiente utilizando nodos Arduino Uno equipados com rádios compatíveis com o padrão IEEE 802.15.4. Os resultados demonstraram que o método de detecção de *outliers* associado à técnica de *data clustering*, são capazes de fornecer informações mais precisas mesmo na presença de *outliers*.

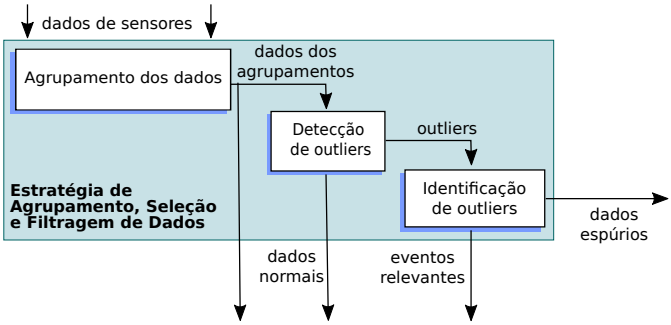


Figura 16 – Estratégia de agrupamento, seleção e filtragem de dados.

Como os nodos são dispositivos tipicamente de baixo custo e alimentados por baterias, *outliers* são eventualmente gerados (HODGE; AUSTIN, 2004), os quais podem introduzir erros que precisam ser tratados adequadamente. Neste tipo de aplicação, dois grandes desafios a serem superados são como lidar com grandes quantidades de dados e, simultaneamente, como reduzir o consumo energético dos nodos.

As RSSFs são formadas, geralmente, por nodos com recursos computacionais limitados mas que têm de lidar com grande quantidade de dados. Nesse contexto, técnicas detecção de *outliers* leves que demandem pouca capacidade computacional são fundamentais. Como visto na Seção 3.3.3, as técnicas estatísticas possuem requisitos promissores para serem usadas na arquitetura proposta.

As técnicas estatísticas são baseadas em baixa exigência de recursos, quando comparadas a outras soluções propostas (baseada em vizinhança e baseada em classificação) (ZHANG; MERATNIA; HAVINGA, 2010; RASSAM; ZAINAL; MAAROF, 2013a). Nas técnicas estatísticas, o modelo de referência é utilizado para analisar a última amostra do dado obtidos antes de decidir sobre a significância, ou não, deste dado.

Essas técnicas podem ser classificadas como paramétricas ou não-paramétricas, e a diferença entre elas encontra-se na presença *a priori* de um modelo de referência. Nas abordagens paramétricas, o modelo de referência é conhecido em tempo de projeto. A abordagem paramétrica é mais simples e assume um modelo gaussiano em que a distribuição de dados subjacente é normal. Por outro lado, na

abordagem não-paramétrica não há conhecimento prévio sobre a distribuição dos dados. Frequentemente, o modelo de referência é construído em tempo de execução através de uma história de execuções anteriores (por exemplo, usando técnicas de regressão do kernel (RASSAM; ZAINAL; MAAROF, 2013a)).

Nesta tese foram analisados cinco métodos estatísticos para a detecção de *outliers*: *critério de Chauvenet* (TAYLOR, 2012), *critério de Peirce* (ROSS; PH, 2003), *Média tolerante a falhas* de Marzullo (MARZULLO, 1990) (FTA), *Confidence-Weighted Averaging* (CWA) do Elmenreich (ELMENREICH, 2007) e a combinação de *Confidence-Weighted Averaging* com *Média tolerante a falhas* (CWA + FTA). Estas foram descritos na Seção 3.3.3.

Complementarmente, para a efetiva detecção dos *outliers*, verificou-se a necessidade de agrupar os valores dos sensores em diferentes *clusters*. Neste sentido, foi aplicada a técnica de *clusterização* de dados *k-means* (RASSAM; ZAINAL; MAAROF, 2013a) e, posteriormente, analisou-se o comportamento dos valores médios obtidos em cada *cluster*. O estudo descrito neste capítulo serviu de base para a escolha dos métodos que foram usados no módulo da estratégia de agrupamento, seleção e filtragem de dados da arquitetura.

5.2 Descrição do Experimento

O experimento foi realizado no Laboratório de Sistemas Embarcados e Robóticos – LASER, na Universidade Federal de Santa Catarina, campus de Blumenau. O experimento consistiu no monitoramento ambiental que usa a temperatura (Celsius)

como grandeza observada.

A Figura 17 mostra a distribuição dos sensores no ambiente, os quais eram fixos e sem mobilidade. Foi utilizada uma rede homogênea de 16 sensores, em uma área de aproximadamente 15m x 10 m. O *hardware* usado está brevemente descrito na Tabela 3.

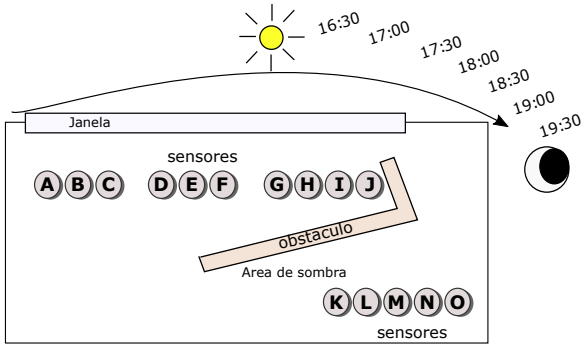


Figura 17 – Cenário de implantação dos Sensores.

A rede foi configurada em uma topologia estrela, tendo seus nodos rádios compatíveis com o padrão IEEE 802.15.4. Um nodo sensor, posicionado sobre a janela, mas não representado na Figura 17, foi configurado para desempenhar o papel de coordenador da rede e sincronizar outros nós, recebendo e armazenando os dados detectados.

A RSSF foi usada para monitorar a temperatura do prédio ao longo do dia. Apesar do monitoramento ter ocorrido em um intervalo contínuo de 24 horas, os valores apresentados neste capítulo representam apenas o período de tempo entre 16:30h e 19:30h. A escolha desse intervalo de tempo é devido à existência de uma maior variação de temperaturas entre os

Tabela 3 – Descrição do Sensor

Hardware	**Função**
Arduino Uno R3	Placa Arduino open-hardware e possui ambiente de desenvolvimento baseado em linguagem C.
Arduino X Bee Shield	Componente desenvolvido para comunicação Wireless do Arduino.
Sensor de Umidade e Temperatura DHT 11	Sensor de temperatura e umidade que permite fazer leituras de temperaturas entre 0 a 50 Celsius e umidade entre 20 a 90%.

sensores. Essa variação foi decorrente do fato de alguns sensores estarem em exposição ao sol naquele intervalo de tempo, enquanto outros permaneceram na sombra. É também importante notar que, embora os sensores tenham operado com uma frequência de uma amostra a cada 30 segundos, por conveniência, para facilitar a compreensão, ao longo deste capítulo, os resultados estão apresentados em intervalos de 30 minutos.

5.3 Avaliação de Técnicas de Detecção de *Outliers* e Fusão da Informação

Experimentos foram projetados com o objetivo de avaliar as técnicas para detecção de *outliers* e fusão de informação. Para tanto, duas configurações foram analisadas: (i) cenário com 10 sensores, sendo que um deles fornece valores que diferem dos outros restantes (Seção 5.3.1); e (ii) cenário com 15 sensores, sendo que os sensores fornecem dados que podem ser classificados e agrupados em agrupamentos de dados

distintos (Seção 5.3.2).

5.3.1 Cenário com 10 sensores

Neste experimento, os dez sensores são representados pelas letras de A a J. O último sensor – sensor J – foi usado neste experimento como uma fonte de *outliers*, devido ao fato da maior parte dos valores entregues serem discordantes quando comparados com valores dos outros sensores (Figura 17). Isso ocorreu porque, no intervalo de tempo analisado, o sensor J permaneceu a maior parte do tempo na sombra, enquanto os outros ficaram expostos ao sol.

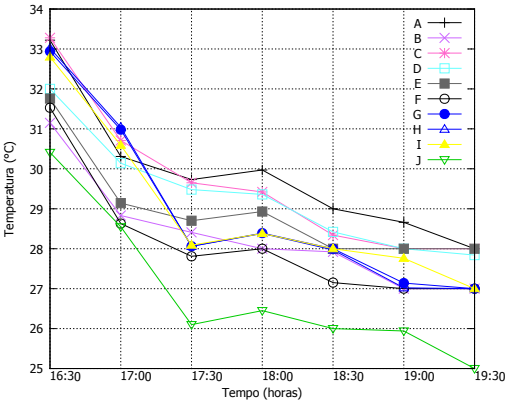


Figura 18 – Valores de temperatura dos sensores.

A Tabela 4 mostra os resultados de detecção de *outliers* para cada técnica avaliada. Os métodos avaliados foram Chauvenet, Peirce, FTA e CWA + FTA. Cada vez que o dado do sensor J foi detectado como um *outlier*, a tabela apresenta a letra "Y"; caso contrário, apresenta a letra "N".

Tabela 4 – Detecção de *outliers* (Y= detectado; N= Não detectado).

Amostras	**Chauvenet**	**Peirce**	**FTA**	**CWA+FTA**
16:30	N	N	Y	Y
17:00	N	N	Y	Y
17:30	Y	Y	Y	Y
18:00	Y	Y	Y	Y
18:30	Y	Y	Y	Y
19:00	Y	Y	Y	Y
19:30	Y	Y	Y	Y

Nos resultados, o critério de Chauvenet teve, em média, um comportamento igual ao Critério de Peirce, foram detectados corretamente *outliers* em apenas 71% das medições. Enquanto, as técnicas FTA e CWA + FTA detectaram corretamente o sensor de J como um *outlier*, durante todo o período monitorado. Por conseguinte, apenas estas duas últimas técnicas serão analisadas na sequência deste capítulo.

5.3.2 Cenário com 15 sensores

Neste experimento foram utilizados dados de quinze sensores nomeados por letras de A a O (Figura 17). Importante notar que o cenário descrito na seção anterior (o cenário com 10 sensores) se constitui em um caso particular deste cenário com 15 sensores. De fato, no cenário anterior, os sensores de K a O foram removidos a partir da análise, permitindo que o sensor J fosse assumido como uma fonte de dados discordantes. Por outro lado, neste novo cenário estudado, o sensor J é acompanhado por outros sensores que também estão próximos ao chão, permanecendo na sombra a maior parte do período analisado.

A Tabela 5 apresenta a média obtida de cada método (em escala Celsius), depois de remover os valores assumidos como *outliers*. Para fins de comparação, a média aritmética simples e o método de Elmereich (CWA) também são apresentados. Em ambos os métodos, não há remoção *outliers*; por outro lado, o método CWA é melhor que a média aritmética simples pois reduz o peso dos de *outliers* nos cálculos da média, ponderando os valores obtidos de acordo com o inverso da sua variância. (A base lógica da técnica CWA é a de assumir que sensores que possuem uma maior variação nos seus dados são menos confiáveis.)

Tabela 5 – Média dos valores das temperaturas usando diferentes técnicas.

Tempo	**Média Simples**	**CWA**	**FTA**	**CWA+FTA**
16:30	31,2 °C	34,2 °C	32,5 °C	34,6 °C
17:00	29,5 °C	30,9 °C	29,5 °C	30,9 °C
17:30	28,1 °C	28,2 °C	28,1 °C	28,2 °C
18:00	27,9 °C	28,2 °C	28,0 °C	28,0 °C
18:30	27,8 °C	28,1 °C	27,8 °C	28,0 °C
19:00	27,5 °C	27,6 °C	28,0 °C	27,2 °C
19:30	27,2 °C	26,8 °C	27,2 °C	27,6 °C

A fim de facilitar a análise do comportamento destes métodos ao longo do tempo, a Figura 19 apresenta os mesmos resultados em um gráfico de linha. É possível observar que, como a média simples utiliza todos os dados para o cálculo, os *outliers* podem prejudicar os resultados. A diferença de valores entre o método FTA e a média simples alcança até 1,3 ^{o} C; e a diferença entre o método CWA + FTA e a média simples chegou até a 3,4 ^{o} C.

Ainda na Tabela 5 são mostradas as diferenças de temperatura entre os métodos, quando comparadas com a

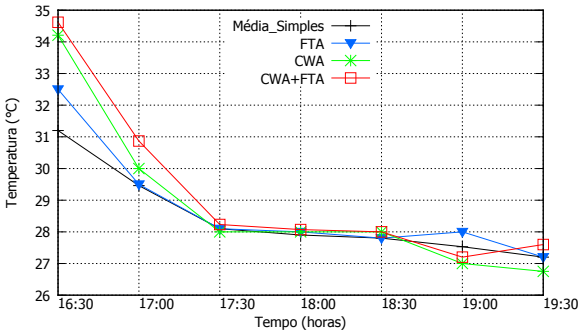


Figura 19 – Comparação das técnicas de fusão da informação.

Tabela 6 – Diferenças dos métodos quando comparados com a média simples.

	CWA		FTA		CWA+FTA	
Tempo	^{o}C	%	^{o}C	%	^{o}C	%
16:30	3,0	9,6	1,3	4,2	3,4	11
17:00	1,4	4,7	0,0	0,1	1,4	4,8
17:30	0,1	0,4	0,0	0,0	0,1	0,5
18:00	0,3	1,2	0,1	0,4	0,2	0,6
18:30	0,3	1,2	0,0	0,0	0,2	0,7
19:00	0,0	0,3	0,5	1,7	0,3	1,2
19:30	0,6	2,4	0,0	0,0	0,4	1,5

média simples, enquanto a Tabela 6 mostra as diferenças entre FTA e método CWA + FTA. Nos cálculos, o método de FTA identifica e remove os *outliers* em todas as amostras, obtendo uma variação de temperatura de até 1,3^{o}C, que representa uma diferença de 4,2%. O método CWA + FTA obtém variação de temperaturas de até 3,4^{o}C, que representa uma diferença de 11%. Comparando FTA com métodos CWA+FTA, a diferença entre temperaturas apresentou variação de até 2,1^{o}C, que

corresponde a 6,8% de diferença, conforme mostrado na Tabela 7.

Tabela 7 – Diferenças entre os métodos FTA and CWA+FTA.

Tempo	**Diferenças**	
16:30	2,1^{o}C	6,8%
17:00	1,4^{o}C	4,7%
17:30	0,1^{o}C	0,5%
18:00	0,1^{o}C	0,3%
18:30	0,2^{o}C	0,7%
19:00	0,8^{o}C	2,9%
19:30	0,4^{o}C	1,5%

5.3.3 Detecção de Eventos em RSSF

No contexto da detecção de eventos, uma importante tarefa no tratamento de *outliers* é a identificação de sua natureza. Como já mencionado, existem diferenças significativas no tratamento de um *outlier* resultante de evento relevante para aquele que é simplesmente um dado espúrio (NAKAMURA; LOUREIRO; FRERY, 2007; RASSAM; ZAINAL; MAAROF, 2013a; KRISHNAMACHARI; IYENGAR, 2003). Nesse contexto, pesquisam-se quão bem abordagens para detecção eventos conseguem identificar a origem do *outlier* pois, em RSSF, a detecção de eventos pode ser utilizada em diversos cenários de aplicações.

Dessa forma, várias técnicas para detecção de eventos são propostas e podem ser classificadas nas seguintes abordagens (PEI et al., 2014), (BAHREPOUR et al., 2009), (YIN; HU; YANG, 2009): baseada em limites, baseada em padrão e baseada em aprendizagem de máquina.

Na abordagem baseada em limites, a detecção ocorre quando a leitura do sensor exceder um valor limite pré-definido havendo uma notificação. A principal vantagem desta abordagem é que os dados podem ser tratados localmente no nodo. Entretanto, valores limites sozinhos são imprecisos e incapazes de capturar características espaço-temporais de eventos, o qual incorre em altas taxas de alarme no monitoramento de aplicações em redes de sensores.

A abordagem baseada em padrões representa eventos espaço-temporais das leituras dos nodos e realiza a detecção de eventos utilizando técnicas de padrões de correspondências. A principal limitação dessa abordagem é a necessidade de padrões de eventos pré-definidos antecipadamente. Portanto, para detecção correta de eventos é necessário que a técnica de padrão de correspondência seja exata para sua aplicação.

Por conseguinte, dentre as técnicas estudadas, a abordagem baseada em aprendizagem de máquina é a única que não precisa de conhecimento prévio das informações aplicando inferências probabilísticas, agrupamentos ou teoria dos grafos para detectar eventos. As abordagens baseadas em limites e as baseadas em padrão necessitam dos valores predefinidos, não sendo adequadas para as RSSF.

Assim, este trabalho foi direcionado para abordagem baseada em aprendizagem de máquina por melhor retratar as condições do cenário de RSSF de larga escala.

5.3.4 Agrupamento de dados através do método *k-means*

As diferenças significativas entre os resultados de diferentes técnicas, conforme mostradas na Tabela 6, podem

ser atribuídas a falsas detecções de *outliers*. A tentativa de se obter uma média única no cenário anterior não é uma abordagem adequada. Torna-se claro que, na aplicação descrita, em determinados períodos de tempo, existem duas médias: a média dos sensores que se encontram no sol e o outra envolvendo os sensores que se encontram na sombra. A tentativa de calcular um valor médio único pode causar, em certas situações, a detecção de *outliers* falsos-positivos.

Nesse sentido, as abordagens de clusterização podem ser utilizadas, permitindo o reunião de sensores, conforme seus valores, dentro de agrupamentos, antes da retirada de *outliers* e cálculo de médias. Existem várias técnicas de clusterização de dados, neste trabalho é aplicada a técnica de *k-means* (ŽALIK, 2008), a qual, na sua forma básica, possui um algoritmo de mineração de dados simples e sem supervisão, e que executa a análise e classifica dados numéricos automaticamente.

Nos experimentos, foi assumida a existência de dois agrupamentos ($k = 2$). Para maior clareza do comportamento da formação dos agrupamentos, na Tabela 8, as letras maiúsculas representam o sensor implantado nas janelas do prédio; e as letras minúsculas representam os sensores implantados próximos ao chão, permanecendo na sombra, na maioria das vezes.

Os dados de temperaturas mais elevadas foram automaticamente agrupados no Agrupamento I; enquanto os dados com temperaturas mais baixas migraram para o Agrupamento II. Pode-se notar que, às 16:30h, todos os sensores nas janelas do prédio ficaram no Agrupamento I e os sensores junto ao chão permaneceram no Agrupamento II. Com

o passar do tempo, observam-se mudanças nos sensores que compõem os agrupamentos.

Tabela 8 – Tabela de transição dos sensores entre agrupamentos.

Tempo	**Agrupamento I**	**Agrupamento II**
16:30	A,B,C,D,E,F,G,H,I,J	k,l,m,n,o
17:00	A,B,C,D,E,G,H,I	F,J, k,l,m,n,o
17:30	A,C,D,E,G,H,I	B,F,J, k,l,m,n,o
18:00	A,C,D,E	B,F,G,H,I,J, k,l,m,n,o
18:30	A,B,C,D,E,F,G,H,I, o	J, k,l,m,n,o
19:00	A,B,C,D,E,G,I, o	F,H,J, k,l,m,n,o
19:30	A,C,D,E, o	B,F,G,H,I,J, k,l,m,n,o

Apesar de terem sido definidos dois grupos, os resultados da Tabela 8 indicam que, a partir das 19:00h, efetivamente temos apenas um único agrupamento. Notadamente, depois das 19:00h alguns sensores trocam aleatoriamente entre os agrupamentos de dados até ao dia seguinte, quando o sol reaparece de novo (período de tempo não mostrado neste trabalho). Essa aleatoriedade foi assumida como uma consequência do fato de, na realidade, haver apenas um único agrupamento de dados, apesar do *k-means* tentar forçar a criação de dois agrupamentos. Por esta razão, a parte restante desta seção discute apenas o intervalo de tempo compreendido entre 16:30h e 18:30h, quando existe claramente a formação de dois grupos.

Como pode ser visto na Tabela 9, com a variação da temperatura ao longo do tempo, os valores médios dos agrupamentos de I e II, apresentam diferenças de até 5,1^{o}C, às 16:30h quando a luz do sol era intensa. Neste sentido, a utilização de clusterização melhora os resultados na detecção

de *outliers* e também a precisão dos valores médios obtidos. A técnica *k-means* calcula uma média simples em cada agrupamento, sem tratar os *outliers*. Por outro lado, como foi referido anteriormente, o método CWA de ponderação de *outliers* reduz os impactos sobre a média, melhorando a precisão.

Tabela 9 – Diferenças entre temperaturas dos agrupamentos usando *k-means*.

Tempo	**Agrupamento I**	**Agrupamento II**	**Diferenças**
16:30	32,9^{o}C	27,8^{o}C	5,1^{o}C
17:00	31,0^{o}C	27,7^{o}C	3,3^{o}C
17:30	29,4^{o}C	27,4^{o}C	2,0^{o}C
18:00	29,3^{o}C	27,5^{o}C	1,8^{o}C
18:30	28,3^{o}C	26,8^{o}C	1,5^{o}C

Nesse sentido, o método CWA também foi aplicado em cada agrupamento definido pelo *k-means*. Os resultados deste método são apresentados na Tabela 10. A diferença das médias obtidas para os diferentes agrupamentos alcançou o valor significativo de até 7,0^{o}C.

Tabela 10 – Diferenças entre os agrupamentos usando o método CWA.

Tempo	**Agrupamento I**	**Agrupamento II**	**Diferenças**
16:30	34,2^{o}C	27,2^{o}C	7,0^{o}C
17:00	31,5^{o}C	29,0^{o}C	2,5^{o}C
17:30	29,4^{o}C	27,2^{o}C	2,1^{o}C
18:00	29,1^{o}C	27,1^{o}C	2,0^{o}C
18:30	28,4^{o}C	26,0^{o}C	2,4^{o}C

A Tabela 11 mostra a diferença entre os valores obtidos utilizando *k-means* com uma média simples e o método CWA

em cada agrupamento. A diferença das médias para os diferentes agrupamentos atingiu o valor de até 1,3^{o}C.

Tabela 11 – Diferenças entre os valores da média simples e o CWA em cada agrupamento.

Tempo	**Agrupamento I**	**Agrupamento II**
16:30	1,3^{o}C	0,6^{o}C
17:00	0,5^{o}C	1,3^{o}C
17:30	0,1^{o}C	0,1^{o}C
18:00	0,2^{o}C	0,3^{o}C
18:30	0,1^{o}C	0,8^{o}C

A técnica de clusterização *k-means* proporcionou um melhor entendimento sobre as especificidades do ambiente monitorado. Tomando como exemplo a amostra das 16:30h, sem a utilização de abordagem baseada em clusterização, usando apenas a média simples, o valor obtido foi 31,2 ^{o}C. Por outro lado, aplicando o método CWA, a média alcançada foi de 34,2^{o}C (Tabela 5). No entanto, para este cenário específico, o correto seria dizer que existem duas médias, como mostrado no resultado clusterizado usando o método CWA (Tabela 10): 34,2^o C para sensores no Agrupamento I (expostos ao sol) e 27,2^{o}C para sensores no Agrupamento II (na sombra).

5.4 Considerações do Capítulo

Aplicações de monitoramento em áreas de grandes dimensões usando RSSF apresentam características importantes, dentre elas, os recursos computacionais limitados dos nodos (capacidade da memória e do processador) e da rede de comunicação (baixa transmissão dos dados). Essas características, associadas à implantação da rede em ambientes

hostis, impactam diretamente na coleta dos dados brutos pelos sensores. Nesta perspectiva, a fusão de dados contribui para maior precisão na coleta dos brutos dados, robustez e na credibilidade dos dados transmitidos.

Contudo, a credibilidade nos dados resultante de alguma técnica de fusão de dados só é alcançada quando esta é associada a alguma técnica de detecção, identificação e tratamento de *outliers*. Neste contexto, o enfoque deste capítulo foi o de avaliar, na arquitetura proposta, os possíveis componentes a serem usados na estratégia de agrupamento, seleção e filtragem de dados. Mais especificamente, foram avaliados métodos leves para detecção e identificação de *outliers* em RSSF.

A vantagem na aplicação das técnicas leves para lidar com *outliers* está no baixo custo computacional, escalabilidade e no fato destas geralmente não necessitarem do histórico anterior dos dados (*stateless*). Tais características contribuem para uma maior eficiência na utilização de recursos de redes limitadas, cenário das RSSF.

Abordagens baseadas em estatística envolvem métodos considerados leves. Dentre os métodos estatísticos avaliados para detecção de *outliers*, os melhores resultados foram o FTA e CWA+ FTA, ambos observaram efetivas retiradas de anomalias e melhor precisão sobre os dados. Por outro lado, os experimentos evidenciaram a necessidade de aperfeiçoar os métodos de detecção de *outliers* através de técnicas de agrupamentos de dados. O uso de agrupamento confere melhor nível de precisão e, consequentemente, uma tomada de decisão com mais exatidão e confiabilidade.

O agrupamento de dados dos sensores utilizando o algoritmo *k-means* permite observar especificidades do ambiente. Para áreas extensas e com diferentes caraterísticas muitas informações podem ser perdidas ou desconsideradas.

Apesar da arquitetura ser voltada para monitoramento em redes de larga escala, devido à dificuldade operacional de se montar um protótipo com grandes dimensões e grande número de nodos usando dispositivos reais, o experimento utilizou apenas 16 nodos em topologia estrela em uma área de cerca de 150 m^2. Embora o protótipo possua pequenas dimensões, os objetivos pré-estabelecidos de se analisar os melhores métodos para a arquitetura foram alcançados.

Observando as análises realizadas, os agrupamentos tiveram diferenças significativas ao longo do monitoramento, justificando o uso de técnicas de agrupamento e evidenciando especificidades que poderiam ser ser mascaradas caso se utilizasse apenas um grupo. Por fim, a utilização de métodos leves para detecção de *outliers* associada ao algoritmo de agrupamento de dados *k-means* conferiu bons resultados para o cenário proposto.

6 Estratégia de Formação de Topologia de Rede Baseada em Dados

Este capítulo descreve a avaliação e resultados da estratégia de formação de topologia em redes *cluster-tree* baseada em dados, da arquitetura proposta, conforme ilustrado na Figura 20. Inicialmente, são discutidos os cenários das simulações e os critérios de avaliação aplicados. Na Seção 6.2, o simulador utilizado e suas características são apresentados. A heurística proposta para formação de topologia baseada em dados é discutida na Seção 6.3. As simulações dos cenários e a análise dos resultados alcançados são descritos nas seções 6.4 e 6.5. No final, a Seção 6.6 apresenta as considerações sobre os resultados obtidos com a estratégia proposta.

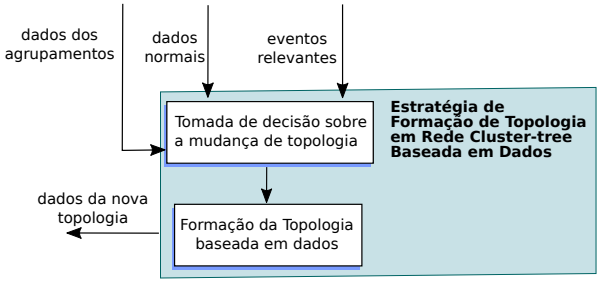


Figura 20 – Estratégia de formação de topologia em redes *cluster-tree* baseada em dados.

6.1 Introdução

Os cenários de RSSFs são muito variados, entretanto uma característica comum é geração de anomalias advindas do nodo (possivelmente por estar descalibrado), da rede (por problemas de conectividade), ou do ambiente (devido a interferências). Uma das premissas deste trabalho foi desconsiderar falhas, interferências e falta de conectividade dos nodos, muito embora estas sejam relevantes[1]. Essa escolha permitiu focarmos especificamente sobre *outliers* de dados, e sobre suas detecções e tratamentos.

O objetivo da estratégia da arquitetura apresentada na seção anterior foi, primeiramente, focar especificamente sobre os dados, reunindo-os em agrupamentos de dados, detectando *outliers* e identificando-os como evento relevante ou dados espúrios. Nesta seção, assume-se que os dados já estão identificados, os agrupamentos formados, e essa informação serve de base para a formação de uma topologia da rede mais adequada para o monitoramento, principalmente com fins de detecção e tratamento de eventos relevantes na área monitorada.

Com esse objetivo, a arquitetura utiliza uma estratégia de formação de topologia em rede *cluster-tree* baseada nos dados de cada nodo, buscando definir as rotas de comunicação desde os nodos sensores até o coordenador. A ideia básica é a

[1] Importante notar que, apesar da arquitetura proposta nesta tese focar especificamente nos *outliers* de dados, outros tipos de anomalias podem ter sido gerados aleatoriamente durante os cenários simulados neste capítulo, pois o simulador utilizado possui modelos realísticos de propagação de rádio, simulando condições de obstáculos, atenuação de sinais etc).

de separar as rotas conforme as características dos dados monitorados na rede. Essas características foram extraídas através dos agrupamentos formados pela técnica *k-means*, e são fornecidas como parâmetros de entrada para uma heurística de formação de árvores *cluster-tree*.

Para melhor descrição da proposta, esta foi aplicada em cenários de RSSF com diferentes características, tais como o tamanho da rede, número de nodos e evento monitorado. Em cada um dos cenários foi observada a evolução do evento detectado, e analisado como a formação da redes baseada em dados pode trazer benefícios na detecção de eventos. Para a análise do desempenho, foram consideradas as seguintes métricas: atraso fim-a-fim, tempo de vida do nodos e taxa de sucesso na entrega dos pacotes.

6.2 Simulador de Rede – Castalia

O simulador selecionado para realizar os testes foi Castalia 3.0[2] desenvolvido na plataforma OMNeT++[3], voltada para RSSF. O OMNeT++ apresenta uma arquitetura de componentes modulares e orientados às aplicações. Estes componentes são desenvolvidos em C++ e permite fácil reutilização, integração e suporte às aplicações.

O Castalia é um simulador com ampla aceitação na comunidade científica devido a sua modularidade, escalabilidade e variabilidade de elementos. Sua arquitetura em camadas permite o desenvolvimento de módulos específicos de protocolos e algoritmos. Tais características possibilitam

[2] https://github.com/boulis/Castalia
[3] https://omnetpp.org

simular cenários com distintas condições e ambientes. A Figura 21 apresenta a estrutura do módulo *nodo* com a disposição das camadas e a suas relações.

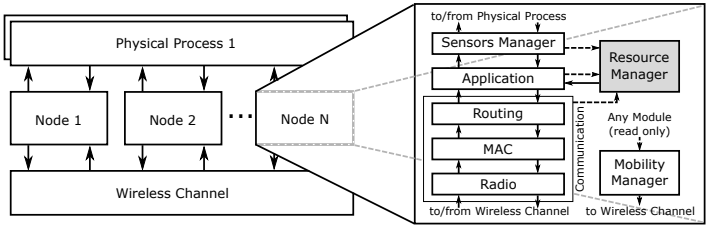


Figura 21 – Estrutura dos módulos e camadas do Castalia. Adaptado de: https://omnetpp.org

Os principais parâmetros de configuração do simulador são aqueles que configuram o rádio, a bateria e parâmetros da comunicação. A Tabela 12 mostra alguns desses parâmetros utilizados para configuração das simulações nos cenários (os parâmetros correspondentes à configuração do IEEE 802.15.4, tais como os que configuram o intervalo de *beacon* e tamanhos do *superframe*, não são mostrados).

Tabela 12 – Parâmetros da configuração das simulações.

Bateria	Estação base	Com alimentação
	Nodos	200 J
Comunicação	Taxa de transmissão	250 Kbps
	Tempo de formação	30 seg
	Tentativas de associação	4
	Pacotes enviados	1000
	Tamanho do buffer do CH	200
Rádio	Cenário Intel Lab Data	- 15 dBm
	Cenários com 100 Nodos	- 1 dBm

Os cenários de simulações que serão apresentados a

seguir seguem essas mesmas configurações da Tabela 12 quanto à bateria e comunicação. Nessa tabela, é possível perceber que a potência sinal do rádio possui parâmetros distintos entre as simulações. Essa questão é motivada pelo tamanho das áreas simuladas e a distância entre os nodos.

6.3 Estratégia de Formação de *Cluster-tree* Baseada em Dados

Após a classificação dos nodos baseada em seus dados e da retirada das anomalias, essas informações são encaminhadas para a estratégia de formação de *cluster-tree*. O objetivo da estratégia é formar rotas dos nodos até o coordenador (estação base), considerando os agrupamentos de dados gerados e, a partir dessas formações, atribuir diferentes taxas de monitoramento para cada uma delas.

No processo de formação da topologia da rede, cada nodo busca estabelecer rotas até o coordenador utilizando o máximo possível de nodos que pertencem ao seu próprio agrupamento. Contudo, quando necessário, um nodo utilizará nodos de outros grupos como uma "ponte" para alcançar o seu destino.

Cada rota formada pode ter uma taxa de monitoramento diferente de outra, em virtude dos dados dos nodos que pertencem a um agrupamento terem características distintas dos dados de nodos que pertençam a outro. Como exemplo, dados de nodos de um agrupamento que estejam monitorando uma região prestes a pegar fogo, serão distintos dos outros dados e tenderão a pertencer ao mesmo

agrupamento. A aplicação da rede poderá configurar a rede para que a taxa de monitoramento das áreas cobertas por esses nodos seja maior que a taxa de monitoramento de nodos das outras áreas. Consequentemente, o atraso fim-a-fim, consumo de energia e taxa de transmissão dos diferentes agrupamentos serão influenciados com o aumento ou diminuição das taxas de monitoramento daquele grupo específico.

A heurística proposta nesta tese é comparada com uma abordagem convencional, a qual é denominada Baseline. Essa abordagem mais simples não considera os dados monitorados para a formação dos *clusters*, e é descrita na Seção 6.3.1. Na sequência, na Seção 6.3.2 é descrita a nossa proposta do algoritmo para formação de *cluster-tree* baseado nos dados.

6.3.1 Algoritmo Baseline

O algoritmo Baseline é descrito e avaliado neste documento apenas para efeitos de comparação com a heurística proposta nesta tese. Esse algoritmo forma uma rede *cluster-tree* considerando apenas, como critério, a conectividade entre os nodos. O processo de formação da rede ocorre em uma abordagem *top-down*, na qual o coordenador (i.e. a estação base) é responsável por iniciar o processo de formação da rede. Para fazer isso, ele assume o papel de um *cluster-head* (CH) e transmite um quadro de convite (o qual é um quadro de enlace específico do protocolo de rede) indicando sua identificação junto com informações básicas (por exemplo, número máximo de nodos em cada *cluster* da rede). Ao receber esse quadro, os nodos enviam quadros de solicitação de associação para o CH. De acordo com critérios específicos, apenas um pequeno

número de nodos pode ser aceito para ingressar no *cluster*. O CH envia mensagens de confirmação para todos os pedidos de associação aceitos e confirmações negativas para os outros. Ele também seleciona e reconhece vários nodos que preencherão o papel de CHs filhos. Esse processo é feito recursivamente, até que todos os nodos estejam associados à rede.

A política de escolha de CH pode ser orientada através de alguns parâmetros específicos. Por exemplo, o número máximo de CHs filhos pode ser limitado para aumentar a profundidade da árvore gerada. Para a implementação específica do Baseline, a escolha dos CHs é feita assumindo que os nodos mais distantes são os melhores candidatos e com número máximo de filhos por CH. A justificativa é aumentar a cobertura máxima da árvore de *cluster*, mantendo um nível aceitável de qualidade de transmissão entre CHs pai e filho.

Os resultados da abordagem Baseline são apresentados nas seções 6.4.1 e 6.5.1.

6.3.2 Algoritmo DbCTF

O DbCTF (*Data Based Cluster Tree Formation*) é a heurística proposta nesta tese, a qual estende a abordagem Baseline com o objetivo de formar rotas entre os sensores até a estação base ou CH compostas exclusivamente por nodos que monitoram valores de dados semelhantes. Ou seja, o objetivo é formar *clusters* da *cluster-tree* e rotas que interligam estes *clusters* até a estação base formadas por nodos do mesmo agrupamento de dados, o qual foi criado a partir da execução *k-means*.

O pseudocódigo DbCTF é descrito no Algoritmo 1.

Por uma questão de simplicidade, a heurística proposta considera apenas dois tipos de grupos de dados: azul e vermelho, onde os nodos vermelhos são aqueles em áreas de monitoramento consideradas de eventos relevantes. O algoritmo DbCTF também inicia a partir do coordenador PAN que seleciona um número X de nodos filhos (linhas 8 e 10). Se houver a associação de nodos vermelhos, o coordenador PAN poderá selecionar pelo menos um desses nodos para garantir uma ramificação composta de nodos que monitoram eventos relevantes.

Basicamente, cada CH cria seu próprio *cluster*, selecionando um número Y de nodos. Os CHs azuis aceitam apenas solicitações de associação de nodos azuis (linha 14). Por outro lado, os CHs vermelhos podem eventualmente aceitar um pedido de associação de um nodo azul (linha 19). Depois de aceitar novos nodos, cada CH seleciona até Z nodos "mais distantes" como seu CH filho. Enquanto CHs azuis selecionam apenas nodos azuis como CH filho, os CHs vermelhos devem selecionar o nodo azul associado como um de seus CHs filhos e defini-lo como CH branco. Os CHs brancos são nodos temporários especiais responsáveis por encontrar nodos vermelhos isolados, até uma profundidade máxima predefinida. Se nodos vermelhos não forem encontrados até uma profundidade máxima, esse conjunto de nodos brancos será desalocado (linhas 29–39). Este procedimento de formação é repetido até que todos os nodos tenham sido associados.

A estratégia DbCTF utiliza os dados para formação da topologia e algumas constantes são norteadoras neste processo. As seguintes constantes foram usadas: X = 3, Y = 6 e Z = 3. De

acordo com esta configuração, a estação base começa com 3 CHs filhos (sendo um deles vermelho); cada *cluster* tem no máximo 6 nodos, além de seus CHs; cada CH tem no máximo 3 CHs filhos. É possível observar que, com a estratégia DbCTF, cada *cluster* é sempre composto por nodos da mesma cor. Além disso, sempre que possível, um caminho para a estação base é formado usando CHs da mesma cor.

Nas seções as seguir, os resultados das simulações são apresentados e avaliados. Foram definidos dois cenários para simulação. O primeiro, na Seção 6.4, usa dados obtidos do projeto Intel Lab Data. O segundo, na Seção 6.5, usa dados gerados em cenários aleatórios, os quais, neste trabalho, serão denominados como cenários estendidos.

Algorithm 1: Heurística DbCTF

```
1  Coordenador PAN assume o papel de CH;
2  repeat
3      foreach CH do
4          CH faz broadcasts com quadro de convite
5          CH recebe solicitações de associação dos nodos;
6          if CH == PAN coordinador then
7              if associando nodo_vermelho > 0 then
8                  CH seleciona até X nodes como seu CH filho
                    sendo pelo menos um nodo vermelho;
9              else
10                 CH seleciona até X nodos azuis como CH filho;
11             end
12         else
13             if tipo_agrupamento(CH) == azul then
14                 CH faz associação de até Y nodos azuis;
15                 CH seleciona até Z nodos como CH filho;
16             end
17             if tipo_agrupamento(CH) == vermelho then
18                 if associando nodo_azul > 0 then
19                     CH faz associação Y − 1 nodos vermelhos e
                        também nodos azuis;
20                     CH seleciona até Z − 1 nodos vermelhos e o
                        nodo azul como CH filho;
21                     CH define nodo azul como grupo de dados
                        branco;
22                     branco profundidade = 0;
23                 else
24                     CH faz associação até Y nodos vermelhos;
25                     CH faz associação ate Z nodos como CH filho;
26                 end
27             end
28             if tipo_agrupamento(CH) == branco then
29                 if enquanto profundidade <
                    max branco profundidade then
30                     if associando nodo_vermelho > 0 then
31                         CH faz associação de um vermelho e
                            seleciona como CH filho;
32                     else
33                         branco profundidade += 1;
34                         CH faz associação de um azul e seleciona
                            como CH filho;
35                         CH define o nodo azul para grupo branco;
36                     end
37                 else
38                     todos os nodos brancos são desalocados;
39                 end
40             end
41         end
42     end
43 until todos os nodos estarem associados;
```

6.4 Análise dos Resultados – Cenário Intel Lab Data

O conjunto de dados utilizados para as simulações e validações da abordagem proposta foram obtidos através do projeto Intel Lab Data[4], conforme mostrado na Figura 22. Os dados monitorados foram coletados em 3 março de 2004. A área utilizada possui 40x35 metros com 54 nodos. Os nodos 5, 15 e 31 foram removidos dos nossos experimentos por apresentarem mau funcionamento durante o teste original, e por possuírem poucos valores armazenados nos *logs*. Para as simulações foram utilizados 30 minutos dos dados coletados em cada um dos horários: 10 e 16 horas. A escolha destes horários foi por apresentarem maiores alterações nos agrupamentos após utilização da técnica *k-means*.

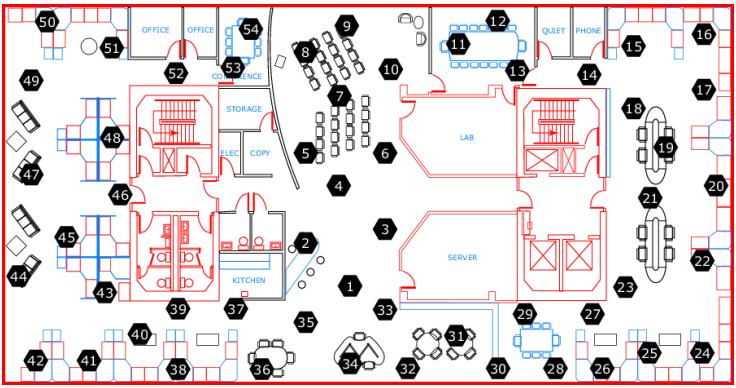


Figura 22 – Intel Lab Data da Intel Berkeley Research lab.

Durante a formação da topologia *cluster-tree*, a informação se um nodo pode estabelecer comunicação com seus vizinhos é um parâmetro importante. Nesse caso, uma matriz

[4] http://db.csail.mit.edu/labdata/labdata.html

de adjacente foi usada, a qual é uma matriz quadrada de ordem N, onde N é o número de nodos da rede. Esse tipo de matriz é composta de valores 1 (indicando que existe conectividade entre o nodo correspondente na linha e coluna da matriz) e 0 (caso contrário).

Como os *logs* do Intel Lab Data não proveem qualquer valor de qualidade do sinal entre os nodos como (ex. valores de RSSI), foram utilizados, com o mesmo objetivo, os valores de taxa de sucesso da comunicação entre os nodos, providos por esta base. No entanto, é assumido neste trabalho que um nodo que tenha taxa de sucesso na comunicação com outro nodo menor ou igual a 20% não tem conectividade. Essa decisão se deve porque é ineficiente assumir como vizinho um nodo que tenha uma baixa taxa de sucesso na comunicação direta. Assim sendo, nessa situação, a posição correspondente na matriz de adjacência recebe o valor 0.

Por outro lado, a matriz de adjacência usada neste trabalho é um pouco diferente da tradicional, porque quando um nodo é capaz de comunicar com o vizinho, em vez de ser armazenado o valor 1 na matriz, o próprio valor da taxa de sucesso é armazenado. Esse valor será importante para o posterior uso do algoritmo de formação do *cluster-tree*, porque na abordagem adotada, um nodo sempre tenta conectar ao CH mais longe possível. Dessa forma, um nodo sempre tentará se conectar com outro nodo da sua matriz de adjacência que possua o valor mais baixo, desde que esse valor seja maior que zero.

Para simular a RSSF, o *framework* Castalia 3.3[5] foi

[5] https://github.com/boulis/Castalia

utilizado junto com a ferramenta CT-SIM (LEÃO et al., 2017). Essa ferramenta provê procedimentos adequados para construir a topologia *cluster-tree*, incluindo o uso de algoritmos de escalonamento de *beacon* adequados para sincronizar a operação com múltiplos *clusters*.

Na abordagem Baseline, o nodo monitora o ambiente com uma taxa de 1 pacote a cada 20 segundos. Por outro lado, na abordagem DbCTF assume-se que existem duas taxas de transmissão dos dados. Os nodos vermelhos, da mesma forma que os nodos da abordagem Baseline, monitoram o ambiente com uma taxa de 1 pacote a cada 20 segundos. Como os nodos azuis monitoram as áreas com temperatura mais baixa, presume-se que eles possam monitorar o ambiente a uma taxa de 1 pacote a cada 100 segundos. A lógica é que os nodos que monitoram regiões com temperatura mais alta (nodos vermelhos) precisam trabalhar com maior frequência do que outros (nodos azuis), já que essas regiões estão sujeitas a uma maior ocorrência de eventos relevantes (por exemplo, incêndios).

Como a topologia Baseline não faz distinção entre os dois tipos de dados, supõe-se que utiliza a taxa mais alta para monitorar o ambiente. A Tabela 13 mostra alguns dos principais parâmetros utilizados para avaliação das simulações no cenário do Intel Lab Data (os parâmetros correspondentes à configuração IEEE 802.15.4 não são mostrados).

A seguir são apresentados os resultados das simulações aplicando os parâmetros da Tabela 13 ao cenário Intel Lab Data. Inicialmente, um cenário com heurística Baseline foi configurado de forma que toda a rede operasse nas mesma taxa

Tabela 13 – Parâmetros da simulação.

Taxa de transmissão de nodos no Baseline e dos nodos do agrupamento vermelho no DbCTF	1 pct a cada 20 s
Taxa de transmissão dos nodos do agrupamento azul no DbCTF	1 pct a cada 100 s
Tempo de simulação	996 min

de transmissão, ou seja 1 pacote a cada 20 segundos. Na sequência, duas configurações foram utilizadas usando os horários de 10 e 16 horas do Intel Lab Data. Nessas configurações, diferentes taxas de transmissão foram atribuídas aos agrupamentos dos dados gerados pelo algoritmo *k-means*.

6.4.1 Cenário com Heurística Baseline

Nesse cenário de simulação, todos os nodos pertencem a um mesmo grupo. Essa é uma configuração usual em RSSF, onde geralmente os nodos não têm diferentes taxas de transmissão.

A Figura 23 apresenta a topologia formada com os nodos na mesma taxa de transmissão utilizando o CT-SIM no cenário Intel Lab Data. O nodo 4 do cenário original do Intel Lab Data (Figura 22) foi assumido como o nodo 0, representando a estação base (coordenador) da *cluster-tree*, por se localizar no centro da área monitorada. Conforme discutido na Seção 6.3.1, todos os nodos possuem conectividade elegendo seu CH e transmitindo até a estação base (nodo 0). Cabe ressaltar que os nodos tentam se associar ao CH mais distante, mas que possua conectividade aceitável.

A Tabela 14 mostra as médias aritméticas dos nodos obtidas com o monitoramento da temperatura e umidade. Nessa

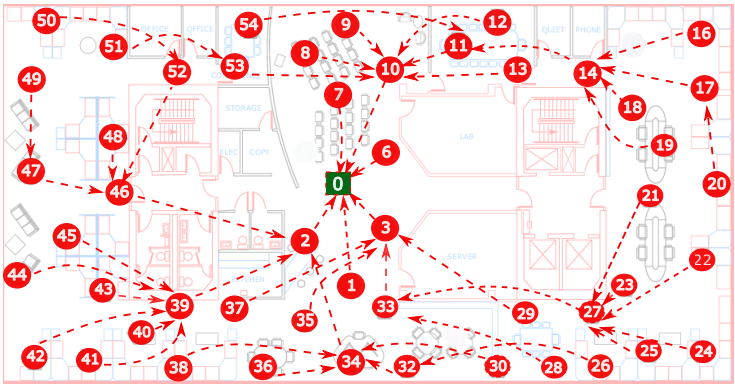


Figura 23 – Formação da topologia Baseline.

rede, o horário utilizado como referência foi o das 10 horas.

Tabela 14 – Valores médios obtidos quando aplicada a heurística Baseline.

	Usando todos os valores
Temperatura	25,66^{o}C
Umidade	25,31 Rh

A Tabela 15 apresenta as métricas de desempenho coletadas na simulação, com relação aos valores médios de atrasos fim-a-fim, tempos de vida e taxas de sucesso.

Tabela 15 – Resultados da simulação com heurística Baseline.

Atraso fim-a-fim	21,85 s
Tempo de vida dos nodos	10,67 h
Taxa de sucesso	34,04 %

6.4.2 Cenário das 10 horas com Heurística DbCTF

Nesse cenário de simulação, os dados coletados são classificados em grupos utilizando a técnica *k-means*. O número de grupos k pode ser determinado conforme necessidade do cenário. Em nossas simulações foi definido $k = 2$, e os dois grupos formados foram denominados agrupamento azul e agrupamento vermelho. A Figura 24 apresenta a posição dos sensores e a relação temperatura/umidade após a utilização do *k-means*.

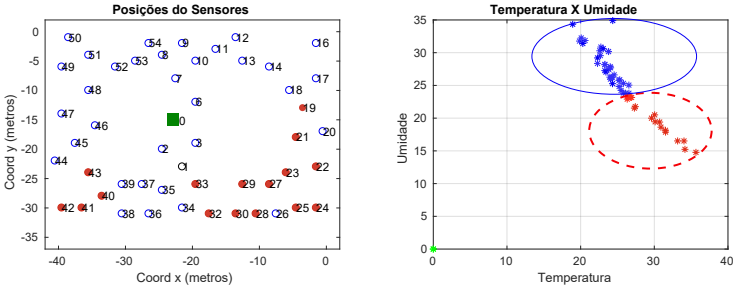


Figura 24 – Intel Lab Data grupos de dados formados pela técnica *k-means* usando dados de temperatura e umidade.

Com a definição dos grupos formados, a Tabela 16 mostra a comparação entre as médias obtidas por todos os nodos contrastando esses resultados com os de valores individuais de cada agrupamento. As diferenças das médias de temperatura entre os agrupamentos é de 5,53^{o}C, e com relação à umidade, a diferença é de 6,23 Rh.

Em tal situação, como citado em (ANDRADE et al., 2016), o uso de técnicas de clusterização de dados é muito importante. É possível analisar, através desses resultados, que o

uso de uma média única de valores dos sensores é inadequado porque, de fato, existem dois grupos de valores, portanto, existem duas médias distintas.

Tabela 16 – Valores das médias usando todos os dados dos nodos e as médias usando os valores individuais dos agrupamentos formados pelo *k-means*.

	Todos os valores	**Agrup. Azul**	**Agrup. Vermelho**
Temperatura	25,66^{o}C	23,60^{o}C	29,13^{o}C
Umidade	25,31 Rh	27,14 Rh	20,91 Rh

Aplicando os mesmos agrupamentos que foram formados e estão ilustrados na Figura 24 ao algoritmo DbCTF é criada a topologia de redes que está ilustrada na Figura 25. Na topologia formada, os nodos que pertencem ao agrupamento vermelho estão representados por círculos hachurados, enquanto os demais, com exceção do nodo 0 que é a estação base, pertencem ao agrupamento azul.

Nas simulações, quando houve formação de ilhas de nodos desconectados, como no caso dos nodos 40, 41, 42 e 43, foram utilizados nodos de outros agrupamentos para formar o caminho até a estação base. Isso é feito porque é importante manter a conectividade de todos os nodos da rede. Entretanto, ao fazer parte da rota de um *cluster* formado por nodos vermelhos, os nodos 35 e 2, que são do agrupamento azul, passam a executar com as mesmas caraterísticas e mesmas taxas de amostragem da rede vermelha (1 pacote a cada 20 s).

Um processo importante de destacar ocorreu com nodo CH 35 durante a formação da rede. Este processo, denominado desassociação, ocorreu a partir do momento que esse nodo passou a fazer parte do caminho de nodos vermelhos, que

possuem característica mais prioritária. Quando um CH azul passa a possuir como filhos nodos vermelhos, ele precisa liberar seus filhos de característica azul e dar exclusividade ao de característica vermelha. Durante a formação da rede, os nodos 34, 36 e 38 utilizavam o 35 como CH. Ao associar o nodo 40 (vermelho) ao nodo 35 (azul), os nodos 34, 36 e 38 são obrigados a encontrar outro CH azul, observando o critério do mais distante. Nesse caso, o nodo 34 e 38 foram associados ao CH 39; o nodo 36 foi adicionado ao CH 1.

Outra funcionalidade importante a destacar do DbCTF refere-se ao tratamento de nodos órfãos. Chamamos de nodos órfãos um nodo azul que tenha de passar por nodos vermelhos para alcançar a estação base, como, por exemplo, o nodo 26. Este nodo com característica azul irá se associar à rede vermelha passando a operar nos parâmetros dessa rede. Esse critério foi definido por ser melhor um nodo operando em taxa de transmissão mais alta, do que um isolado e sem transmitir dados.

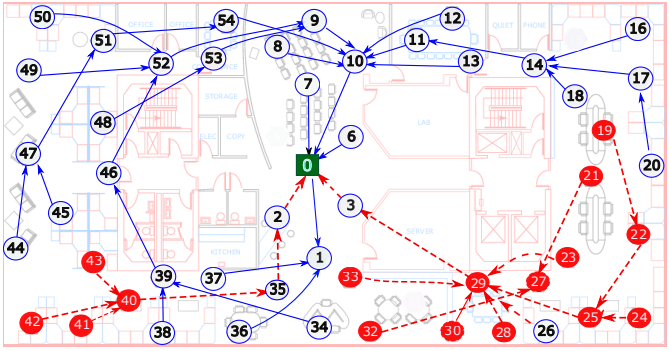


Figura 25 – Topologia da rede no cenário do Intel Lab Data às 10h.

Tabela 17 – Resultados da simulação no Intel Lab Data às 10h.

	Baseline	**Agrupamento Azul**	**Agrupamento Vermelho**
Atraso fim-a-fim	21,85 s	43,81 s	19,84 s
Tempo de vida dos nodos	10,67 h	23,89 h	10,10 h
Taxa de sucesso	34,04%	48,21%	39,16%

Os resultados da simulação com suas métricas são apresentados na Tabela 17. Como resultado da simulação do DbCTF, observa-se uma redução do atraso fim-a-fim dos nodos dos agrupamentos vermelho de 9,20% quando comparados à abordagem Baseline. Essa é uma característica importante quando a rede é instalada com objetivo de detectar eventos relevantes e em regiões da rede que possam precisar de mais prioridade. No DbCTF os nodos do agrupamento azul não precisam monitorar a rede com taxa máxima, assim esses nodos podem ter o benefício de reduzir suas taxas de transmissão para 1 pacote a cada 100 segundos, permitindo aumentar seus tempos de vida. Além disso, todos os nodos da rede podem se beneficiar desta característica, devido à redução do uso compartilhado do meio sem fio.

6.4.3 Cenário das 10 horas com Heurística DbCTF - Com tolerância a faltas

Nesse cenário de simulação, os dados coletados também foram classificados em grupos utilizando o técnicas *k-means* e o número de grupos k foi definido como $k = 2$. Aplicando os mesmos agrupamentos que foram formados e estão ilustrados na Figura 24 ao algoritmo DbCTF é criada a topologia de redes que está ilustrada na Figura 26.

Na topologia formada, os nodos que pertencem ao

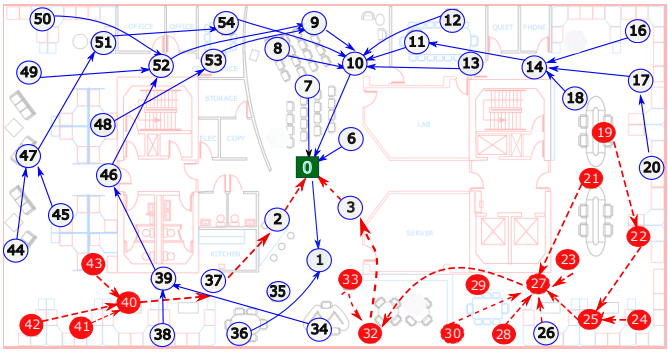


Figura 26 – Topologia da rede no cenário do Intel Lab Data às 10h.

agrupamento vermelho estão representados por círculos hachurados, enquanto os demais, com exceção do nodo 0 que é a estação base, pertencem ao agrupamento azul. Neste processo novos nodos são selecionados com pontes para a rede vermelha, são os nodos 32 e 37. O nodo 26 azul se associa ao CH 27 vermelho, antes associado ao nodo 29. Este nodo com característica azul irá se associar à rede vermelha passando a operar nos parâmetros dessa rede.

Tabela 18 – Resultados da simulação no Intel Lab Data às 10h - com tolerância a faltas.

	Baseline	**Agrupamento Azul**	**Agrupamento Vermelho**
Atraso fim-a-fim	21,85 s	44,74 s	19,55 s
Tempo de vida dos nodos	10,67 h	24,56 h	10,21 h
Taxa de sucesso	34,04%	49,51%	41,37%

Os resultados da simulação com suas métricas são apresentados na Tabela 18. Como resultado da simulação do DbCTF, observa-se uma redução do atraso fim-a-fim dos nodos

dos agrupamentos vermelho de 10,5% quando comparados à abordagem Baseline.

Um aspecto relevante apresentado é a tolerância a faltas, quando a rede vermelha seleciona outros nodos para fazerem parte da rota até a estação base. Essa é uma característica importante quando a rede é instalada com objetivo de detectar eventos relevantes e em regiões da rede que possam precisar de mais prioridade.

6.4.4 Cenário das 16 horas com Heurística DbCTF

Nesse cenário de simulação, os dados coletados também foram classificados em grupos utilizando o técnicas *k-means*, e o número de grupos k foi definido como $k = 2$. A Figura 27 apresenta a formação dos grupos e suas posições, também a relação temperatura e umidade.

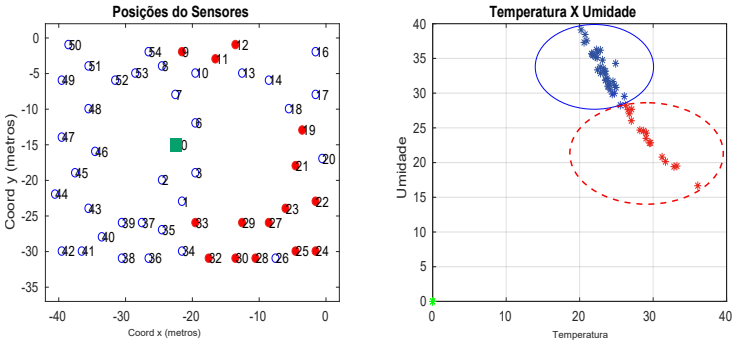


Figura 27 – Intel Lab Data grupos de dados formados pela técnica *k-means* usando dados de temperatura e umidade.

Com a definição dos grupos formados, a Tabela 19

mostra a comparação entre as médias obtidas com todos os dados dos nodos contra os valores individuais de cada agrupamento. A diferença entre as temperaturas médias dos agrupamentos alcançou 6,4^{o}C e entre as umidades médias foi de 9,98 Rh.

Tabela 19 – Comparação da média usando todos os valores contra os valores individuais dos agrupamentos formados pelo *k-means*.

	Todos os valores	**Agrup. Azul**	**Agrup. Vermelho**
Temperatura	25,10^{o}C	23,18^{o}C	29,58^{o}C
Umidade	30,42 Rh	33,41 Rh	23,43 Rh

Aplicando os agrupamentos da Figura 27 ao algoritmo DbCTF observa-se uma formação de topologia baseada nos dados dos agrupamentos, conforme ilustrado na Figura 28. Nessa simulação, os nodos vermelhos 9, 11 e 12 estão isolados e utilizam o CH 10 que é azul como uma ponte para estabelecer uma conexão com a estação base. A característica do CH 10, antes azul, passa a ser vermelha, e todos os nodos azuis associados ao CH 10 são desassociados aos outros CH azuis. Além disso, outra questão importante de se destacar é a associação do nodo 26 azul órfão à rede vermelha, passando a operar uma taxa de transmissão mais alta.

Os resultados da simulação com suas métricas são apresentados na Tabela 20.

Tabela 20 – Resultados da simulação no Intel Lab Data às 16h.

	Baseline	**Agrupamento Azul**	**Agrupamento Vermelho**
Atraso fim-a-fim	21,85 s	41,64 s	17,41 s
Tempo de vida dos nodos	10,67 h	25,41 h	10,11 h
Taxa de sucesso	34,04%	53,01%	31,39%

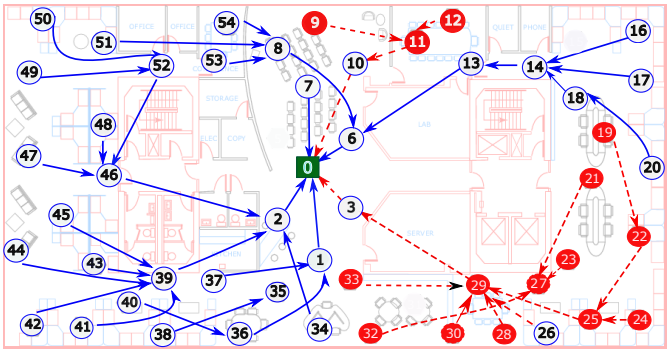


Figura 28 – Topologia da rede no cenário do Intel Lab Data às 16h.

Como resultado da simulação do DbCTF, observa-se a redução de 20,32% do atraso fim-a-fim do agrupamento vermelho, quando comparado à abordagem Baseline. Uma melhora significativa no atraso é muito relevante em regiões que necessitem maior prioridade. Com relação à taxa de sucesso, há uma pequena queda, supostamente devido à formação dos agrupamentos de dados em função da topologia.

Após as simulações realizadas com o cenário de Intel Lab Data nos horários de 10 e 16 horas, foi possível constatar os resultados relevantes gerados pelo DbCTF. A formação topológica baseada nos agrupamentos formados pelos próprios dados monitorados aponta benefícios para rede, especificamente nos tempos de respostas, no tempo de vida do nodos e na redução do acesso ao meio compartilhado.

6.5 Análise dos Resultados – Cenários com 100 nodos

No processo de análise do comportamento do algoritmo DbCTF, também foram gerados cenários em que, de forma aleatória, alguns parâmetros da simulação eram alterados no início de cada simulação, através de recursos da ferramenta Castalia. Diversas simulações com variados números de nodos e posicionamentos foram executadas, e algumas dessas simulações foram selecionadas e são mostradas a seguir. Em comum a todas as simulações é que o cenário envolve RSSF de 100 nodos, além de uma estação base colocada exatamente no centro de uma área de 150X150 metros.

Ao todo, são apresentados a seguir cinco cenários. Inicialmente, um cenário composto totalmente por nodos que se comportam como se um evento relevante estivesse ocorrendo em toda a rede – Cenário de Rede Vermelha. Na sequência é mostrado o comportamento do algoritmo em um cenário oposto do anterior, onde a rede se comporta como se não fosse monitorar eventos relevantes – Cenário de Rede Azul.

6.5.1 Cenário de Rede Vermelha

A Figura 29, apresenta a topologia formada com os nodos na mesma taxa de transmissão utilizando a abordagem do CT-SIM. Nesta simulação, que é equivalente ao cenário de execução do algoritmo Baseline (1 pacote a cada 20s), todos possuem conectividade elegendo seus CHs e transmitindo a estação base (nodo 0). Como critério para se associar os nodos buscam os CHs com conectividade aceitável, acima de 20% de taxa de sucesso na transmissão.

Os parâmetros da simulação são os mesmos apresentados na Tabela 13.

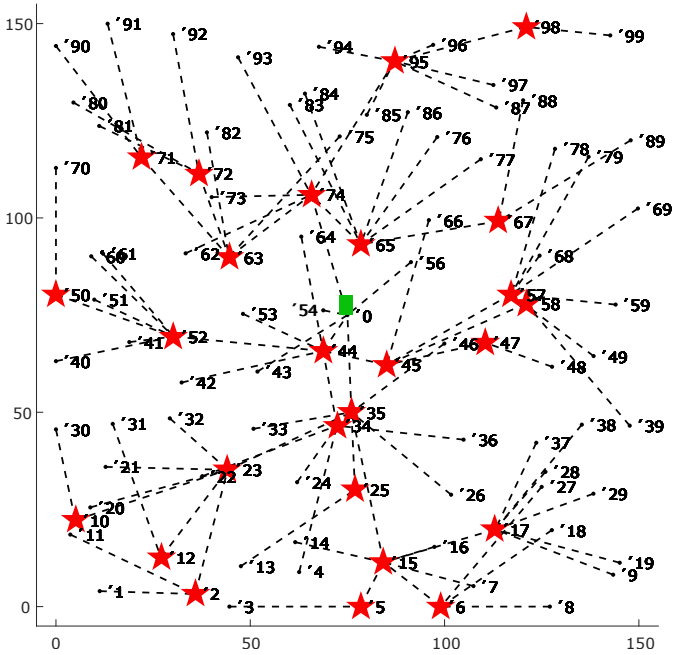


Figura 29 – Topologia da rede no Cenário de Rede Vermelha.

Após a simulação, o desempenho do algoritmo foi avaliado, e a Tabela 21 apresenta o comportamento que foi utilizado para comparação com os resultados dos outros cenários.

6.5.2 Cenário de Rede Azul

Nesta simulação foi adotada a característica da redes azul para toda a rede. Nosso objetivo foi observar o

Tabela 21 – Resultado da simulação no Cenário de Rede Vermelha.

Atraso fim-a-fim	26,79 s
Tempo de vida dos nodos	10,40 h
Taxa de sucesso	15,94%

comportamento do algoritmo DbCTF na formação das rotas em função dos dados.

A Figura 30, apresenta a topologia formada com os nodos na mesma taxa de transmissão utilizando a abordagem do CT-SIM. Nesta simulação, é enviado 1 pacote a cada 100s e todos os nodos possuem conectividade elegendo seus CHs e transmitindo para a estação base (nodo 0). Como critério para se associar os nodos buscam os CHs com conectividade aceitável, acima de 20% de taxa de sucesso na transmissão.

Durante a simulação, o desempenho do algoritmo foi estudado segundo as métricas preestabelecidas, e a Tabela 22 apresenta o os resultados.

Tabela 22 – Resultado da simulação no cenário de Rede Azul.

Atraso fim-a-fim	33,50 s
Tempo de vida dos nodos	26,07 h
Taxa de sucesso	35,20%

Como resultado da simulação do DbCTF contata-se a aumento do atraso fim-a-fim da rede azul quando comparado a abordagem Baseline. O tempo de vida do nodos aumentado, devido ao maior intervalo sensoriamento. Quanto a formação da redes, apresentou a mesma topologia da rede Baseline.

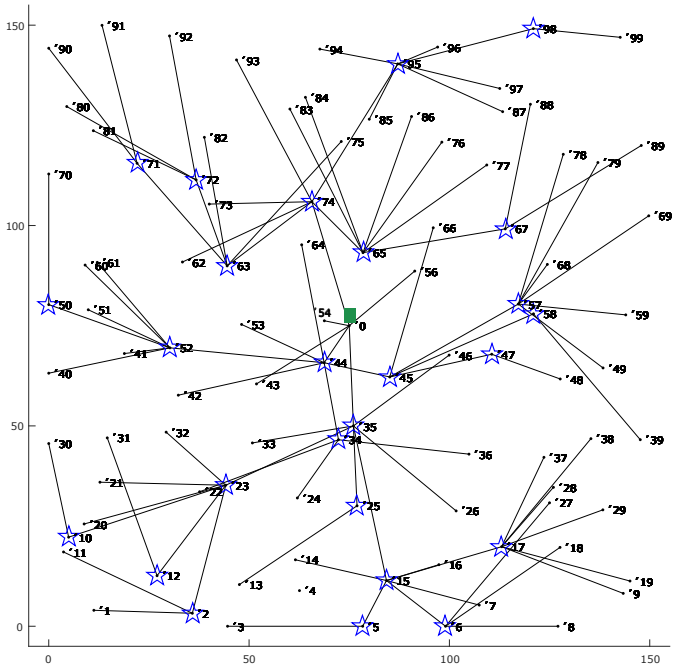


Figura 30 – Topologia da rede no Cenário de Rede Azul.

6.5.3 Cenário de Rede I

Neste cenário estendido, não foi utilizado conjunto de dados de cenários existente para aplicar a técnica *k-means*. O principal objetivo é observar o comportamento do algoritmo Db-CTF na formação das rotas em função dos dados. Uma vez que a justificativa para utilizar agrupamentos de dados e seus resultados já foram discutidos em (ANDRADE et al., 2016).

Para escolha dos nodos que fazem parte do agrupamento azul e agrupamento vermelho, optou-se por criar um evento na rede determinando os nodos que fazem parte. Neste caso, os

nodos vermelhos serão os: 1, 2, 3, 4, 10, 11, 12, 13, 14, 20, 21, 22, 23 e 24. Os demais nodos fazem parte do *tipo_agrupamento* azul.

A Figura 31 mostra a aplicação do DbCTF na formação da topologia em função dos dados. Neste cenário observa-se que o agrupamento vermelho não tem ligação direta com a estação base (nodo 0), sendo necessário utilizar o nodo 35 azul como ponte. O nodo 35 anteriormente com taxa de transmissão de 1 pacote a cada 100 segundos passar a ter característica de 1 pacote a cada 20 segundos.

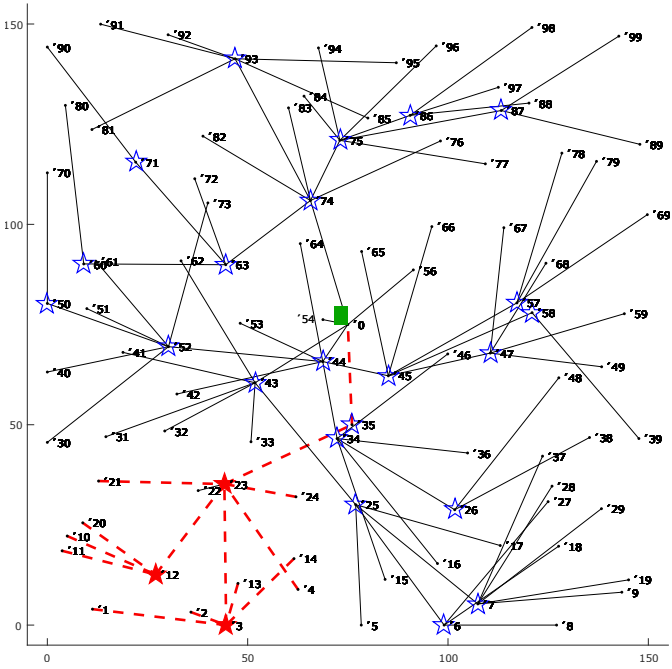


Figura 31 – Topologia da rede no Cenário Redes I.

Na Figura 31 é possível observar o processo de desassociação que o CH 35 realiza liberando os nodos 15, 25, 34, 45, 46 e 53 associados a ele. Estes buscam associação a outros CHs azuis disponíveis. Como o nodo 15 associando-se ao CH 25 azul, os nodos 25 e 46 associando-se ao CH 34 e nodo 53 se associando ao CH 44.

Após a simulação as métricas coletas são apresentadas na Tabela 23.

Tabela 23 – Resultado da simulação no Cenário Rede I.

	Baseline	**Agrupamento Azul**	**Agrupamento Vermelho**
Atraso fim-a-fim	26,79 s	39,34 s	15,34 s
Tempo de vida	10,40 h	25,55 h	10,20 h
Taxa de sucesso	15,94%	44,64%	25,75%

Como resultado da simulação do DbCTF constata-se uma redução de 42,73% no atraso fim-a-fim da rede vermelha, quando comparada com a abordagem Baseline. Fica evidenciado o destaque na região relevante e sua prioridade na rede. O tempo de vida do nodos aumenta significativamente na rede azul. As taxas de sucesso, tanto no agrupamento azul como no agrupamento vermelho, também obtiveram melhorias. Além disso, toda a rede é beneficiada pela redução do acesso ao meio de transmissão.

6.5.4 Cenário de Rede II

Com base na Rede I, nessa simulação foi aumentado o número de nodos vermelhos e inseridos em outra área da rede. O objetivo foi observar o comportamento do algoritmo DbCTF

na formação das rotas em função dos dados.

Para a escolha dos nodos que fazem parte dos agrupamentos azul e vermelho, optou-se por criar um evento na rede determinando os nodos que fazem parte. Nesse caso, os nodos vermelhos foram os seguintes: 1, 2, 3, 4, 10, 11, 12, 13, 14, 20, 21, 22, 23 e 24, além dos nodos 90, 91, 92, 93 e 94. Os demais nodos fizeram parte do agrupamento azul.

A Figura 32 mostra a aplicação do algoritmo DbCTF na formação da topologia em função dos dados. Nessa simulação, observa-se que os nodos vermelhos não tiveram conexão direta com a estação base (nodo 0), tendo sido necessário utilizar nodos azuis como pontes da transmissão. Neste caso, temos os nodos azuis 35 e 74 atuando como pontes para os nodos vermelhos, passando a operar na condição da rede vermelha, com uma taxa de comunicação de 1 pacote a cada 20 segundos.

Na Figura 32, é possível observar o processo de desassociação que o CHs 35 realiza, liberando os nodos associados a ele. Já o nodo 74 passa a ser um CH azul sendo utilizando como ponte para alcançar a estação base.

Após a simulação, os dados de desempenho são apresentados na Tabela 24.

Tabela 24 – Resultado da simulação no Cenário de Rede II.

	Baseline	**Agrupamento Azul**	**Agrupamento Vermelho**
Atraso fim-a-fim	26,79 s	44,12 s	17,92 s
Tempo de vida dos nodos	10,40 h	25,80 h	10,16 h
Taxa de sucesso	15,94%	44,64%	27,18%

Como resultado da simulação do DbCTF, constata-se a redução do atraso fim-a-fim da rede vermelha para 33,10%

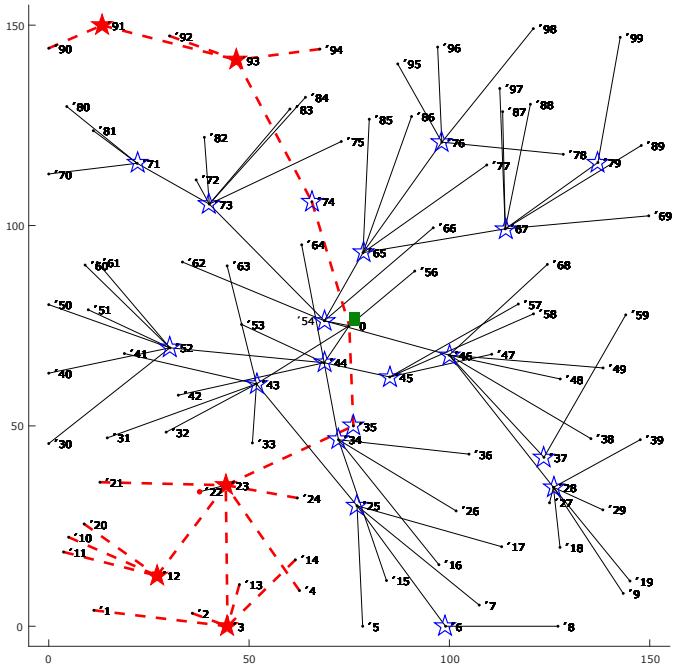


Figura 32 – Topologia da rede no Cenário de Rede II.

quando comparado com a abordagem com Cenário da Rede Vermelha (que, por sua vez, é equivalente ao funcionamento do algoritmo Baseline). Fica evidenciado o destaque na região relevante e sua prioridade na rede. O tempo de vida do nodos aumenta significativamente na rede azul. A taxa de sucesso tanto no agrupamento azul e como no agrupamento vermelho também obteve melhorias. Além disso, importante destacar que toda a rede acaba sendo beneficiada pela redução no acesso ao meio sem fio compartilhado.

6.5.5 Cenário de Rede III

Com base no cenário da Rede II, nessa simulação foi aumentado o número de nodos vermelhos e inseridos em outra área da rede. Nosso objetivo foi observar o comportamento do algoritmo DbCTF na formação das rotas em função dos dados.

Para escolha dos nodos que fizeram parte do agrupamento azul e do agrupamento vermelho, optou-se por criar um evento na rede determinando os nodos que fazem parte desse evento. Nesse caso, os nodos escolhidos como vermelhos foram os numerados de 1 a 29, além dos nodos numerados de 90 a 99. Os demais fizeram parte do agrupamento azul, ou seja, nodos numerados de 30 a 89.

A Figura 33 mostra a aplicação do algoritmo DbCTF na formação da topologia em função dos dados. Nesta simulação observa-se que os nodos vermelhos não tiveram conexão direta com a estação base (nodo 0), tendo sido necessário utilizar nodos azuis como pontes nessa comunicação.

Neste caso, os nodos azuis 32, 33, 34, 35, 36, 37, 38, 40, 41, 44, 45, 47, 49, 50, 52, 59, 65, 66, 71, 73, 74, 75, 77, 83 e 88 são pontes para os nodos vermelhos, passando a operar na condição da rede vermelha de 1 pacote a cada 20 segundo.

A Figura 33 evidencia que, dada a distribuição inicial dos nodos vermelhos e o uso de nodos azuis como pontes para alcançar a estação base, apenas os nodos 43, 54 e 56 mantiveram sua condição inicial de nodos azuis, ou seja, monitoramento com uma taxa de 1 mensagem a cada 100 segundos.

Como resultado da simulação do DbCTF temos que a peculiaridade do evento gerou um condição em que a rede teve

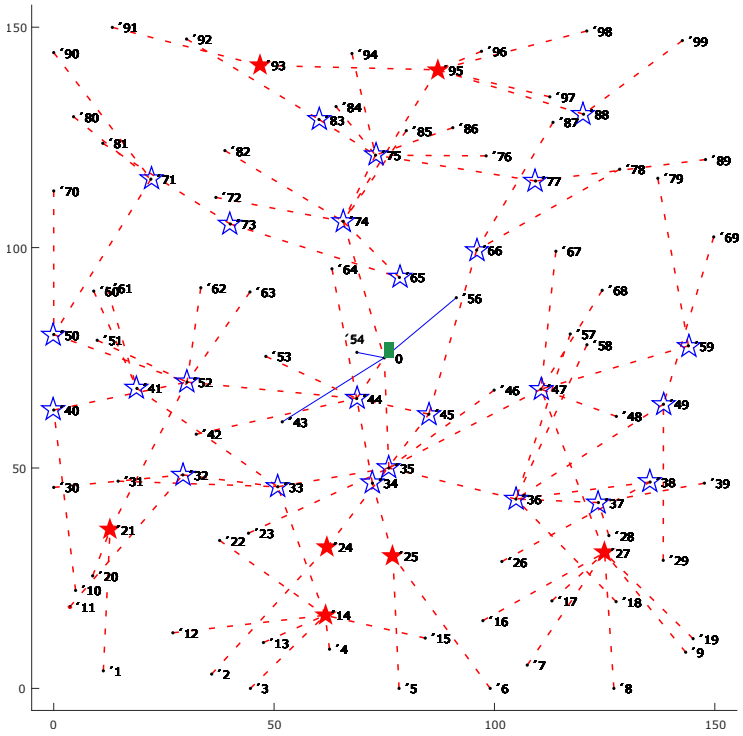


Figura 33 – Topologia da rede no Cenário de Rede III.

sua característica alterada dada a criticidade da redes vermelha em relação a rede azul. Neste cenário, o comportamento da rede se tornou igual a resultado da Tabela 21 com a redes na característica vermelha.

6.6 Considerações do Capítulo

Neste capítulo foi avaliado o algoritmo DbCTF, através de simulações em cenários distintos. No primeiro

conjunto de simulações, a coleção de dados do Intel Lab Data foi usada e observou-se o comportamento nos horário 10 e 16 horas. Os resultados obtidos comprovam a eficiência da proposta.

No segundo conjunto de simulações, foi usado um cenário com 100 nodos em uma área de 150x150 metros. Nesse cenário foi determinada a formação dos nodos que fazem parte do experimento e analisado o comportamento do algoritmo DbCTF. Novamente os resultado obtidos comprovam a eficiência da proposta.

A fim de destacar as contribuição da estratégia de formação em função dos dados, destacamos os seguintes aspectos:

- Tempo de sensoriamento diferenciados por área, resultando em melhor monitoramento do evento;
- Formação de topologia baseada nos dados, trazendo eficiência a rede;
- Melhoria no desempenho do atraso fim-a-fim, fundamental na detecção de eventos relevantes;
- Melhoria no consumo energético da rede, por diferenciar as redes em alta e baixa prioridade.

7 Conclusões da Tese e Trabalhos Futuros

Com potencialidade para uso em vários cenários de aplicação, as RSSFs vêm ganhado grande destaque como área de investigação. As aplicações dessas redes no contexto de monitoramento de grandes áreas são as mais variadas, envolvendo, por exemplo, agricultura de precisão, indústrias, cidades inteligentes e monitoramento ambiental.

Esta tese tratou de assuntos relacionados ao monitoramento de RSSF de larga escala, através de uma proposição de arquitetura voltada para detecção, identificação e tratamento de *outliers*. Como as RSSF necessitam executar de forma autonômica, uma importante contribuição da arquitetura é a proposição de uma estratégia que ajusta a topologia da rede de acordo com os próprios dados que estão sendo monitorados. Dessa forma, consegue-se conciliar os objetivos de operar a rede de forma convencional, através de técnicas de fusão da informação atreladas a técnicas que detectam e tratam os *outliers*, ao mesmo tempo em que se consegue obter um melhor desempenho no monitoramento de eventos relevantes.

7.1 Visão Geral do Trabalho

A pesquisa realizada apresenta uma arquitetura para detecção, identificação e tratamento de outliers em RSSF de larga escala. Complementada, por uma estratégia inovadora na

formação de topologia *cluster-tree* baseada em dados para redes de larga escala.

No Capítulo 2, discutimos os aspectos teóricos das RSSFs sua padronização e seus protocolos de roteamento com ênfase no hierárquico. Cabe ressaltar que a literatura preconiza a topologia *cluster-tree* como a mais indicada para aplicações de RSSF em larga escala. As vantagens na utilização da topologia *cluster-tree* aplicadas em larga escala passaram por sincronização do tempo de operação (uso de *beacon*), consumo energético, comunicação sem colisão e operação com ciclo de trabalho definidos (*duty-cycle*). Em nossa proposta, como mecanismo de comunicação base, foi utilizado o trabalho de (LEÃO et al., 2017). Este desenvolveu um mecanismo eficiente para RSSF de larga escala que não é padronizado pelo IEEE 802.15.4. Sobre esse mecanismo foi desenvolvida nossa estratégia inovadora para formação *cluster-tree* baseada em dados. Outra questão tratada no Capitulo 2 é sobre um elemento essencial para a arquitetura, o conceito de fusão da informação. Estes conceitos são necessários para definir o modelo de fusão da informação aplicado na arquitetura considerando suas potencialidades e as limitações das RSSFs.

O Capítulo 3 discorreu sobre conceitos para detecção de eventos em RSSF de larga escala, integrando seu uso a métodos de aprendizagem de máquina, fusão da informação e tratamento de *outliers*. Esses entendimentos são bases para a arquitetura proposta. Nesses aspectos existem ainda muitas questões em aberto, que motivam o desenvolvimento de novas abordagens para melhorar a eficiência no tratamento da detecção de eventos em RSSF de larga escala. Podemos

evidenciar quanto ao: método aplicado para detecção de eventos, método para identificação/tratamento de *outliers* e o mecanismo para formação da topologia do *cluster-tree* para larga escala. Vários parâmetros podem impactar na seleção dos métodos para detecção de eventos. Nessa direção, as escolha dos métodos podem influenciar no desempenho rede, mais ainda, esses parâmetros podem ter um custo computacional alto sendo incompatíveis com a natureza limitada de recursos da RSSF. Por outro lado, a topologia de rede *cluster-tree* apresenta desafios nos aspectos da comunicação principalmente no fluxo centralizado das transmissões a estação base. Mais especificamente com relação a: consumo energético dos nodos, atraso fim-a-fim, congestionamentos e número de saltos dos nodos ao coordenador. Nesse contexto, esta tese apresentou uma arquitetura para detecção de eventos de RSSF de larga escala associada uma estratégia inovadora para formação da topologia de comunicação *cluster-tree* baseada em dados.

No Capítulo 4, introduzimos a arquitetura para detecção, identificação e tratamento de *outliers* em redes de sensores de larga escala. Esta arquitetura, descrita de forma modular, permite alterações com relação ao método de formação de agrupamentos, método de detecção de eventos e na estratégia de formação da topologia *cluster-tree*. Desse modo, essa arquitetura adequa-se a cenários distintos de RSSF, com a possibilidade de utilizar outros métodos de aprendizagem de máquina para classificar os dados.

No Capítulo 5, apresentou-se os resultados obtidos com a estratégia de agrupamento, seleção e filtragem dos dados. A motivação para o desenvolvimento dessa estratégia

passa pela junção do método de agrupamento com o método para detecção de *outliers* que apontam resultados mais precisos, quando comparados aos mesmos de forma individualmente. Outro aspecto relevante é a modularidade da estratégia possibilitando a utilização de diferentes métodos de agrupamento e de detecção. Importante destacar a relação custos-benefícios no consumo dos recursos computacionais dos métodos escolhidos. A análise dos experimentos comprovam a eficiência da cooperação dos métodos de agrupamento e detecção de *outlier* na precisão dos dados monitorados.

Por fim, no Capítulo 6 apresentou os resultados obtidos com a estratégia de formação de *cluster-tree* baseado em dados. A ideia base é utilizar os dados obtidos no monitorado agrupá-los por intervalos e utilizar os nodos do mesmo grupo, se possível, para alcançar a estação base. Os resultados das simulações utilizando cenários com diversas variações comprovam que abordagem DbCTF foi capaz de atingir seu objetivo principal, reduzir o tempo resposta nas áreas onde podem ocorrer eventos relevantes e aumentar o tempo de vida dos nodos. Além disso, por reduzir o taxa de transmissão em algumas áreas da rede, a taxa de sucesso das mensagens enviadas também aumentou.

7.2 Contribuições da tese

Por fim, as principais contribuições dessa tese podem ser resumidas:

- Em uma arquitetura modular para detecção de eventos em RSSF de larga escala;

- Estratégia inovadora de formação topológica baseada em dados para redes cluster-tree;
- Na melhoria da precisão na detecção do evento para RSSF de larga escala;
- Na adaptabilidade às mudanças e no atendimento online do monitoramento;
- Na tomada de decisão eficiente na detecção, identificação e tratamento de outlier;

7.3 Trabalhos Futuros

A continuidade deste trabalho tem promissoras direções quanto:

- Mecanismo de escolha otimizada do *cluster-head* considerando a melhor cobertura da área monitorada;
- Classificar a reputação dos nodos para seleção das rotas tornando-as mais confiáveis e precisas;
- Mecanismos de disparo automático para reconfiguração da rede observando a alteração dos nodos nos agrupamentos;
- Estender a arquitetura envolvendo outras cenários de aplicação como: Internet das Coisas e *Body Area Network*.

7.4 Lista de Publicações

Os trabalhos produzidos e publicados nesta tese são apresentados a seguir.

Artigos de Conferência

Andrade, A.T.C.; Siedersberger, D.; Montez, C.; Moraes, R.; Leão, E.; Vasques, F. **Data-Based Cluster-Tree Formation Scheme for Large-Scale Wireless Sensor Networks**. In 16th IEEE International Conference on Industrial Informatics (INDIN), 2018. IEEE Conference, July 2018. DOI: 10.1109/INDIN.2018.8471938

Andrade, A.T.C.; Montez, C.; Moraes, R.; Pinto, A.R.; Vasques, F.; Da Silva, G.L.. **Outlier detection using k-means clustering and lightweight methods for Wireless Sensor Networks**. In: IECON 2016 42nd Annual Conference of the IEEE Industrial Electronics Society, October, 2016. DOI: 10.1109/IECON.2016.7794093

Artigo de Periódico

André, P. B.; Andrade, A.T.C. ; Callegaro, R.; Montez, C.; Moraes, R.; Pinto, A. **An Architecture for Information Fusion and for Detection, Identification and Treatment of Outliers in Wireless Sensor Networks**. Communications in Computer and Information Science. 3ed.: Springer International Publishing, 2017, v. 702, p. 81-100. DOI:10.1007/978-3-319-61403-8_5

Referências

AIKENETA. Report of nsf workshop on network research testbeds. *National Science Foundation {...}*, n. November, 2002. 123

AKYILDIZ, I. et al. Wireless sensor networks: a survey. *Computer Networks*, v. 38, n. 4, p. 393–422, 2002. ISSN 13891286. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1389128601003024>. 28, 32, 43, 45, 124

AKYILDIZ, I.; WANG, X.; WANG, W. Wireless Mesh Networks: A Survey. *Computer Networks*, v. 47, p. 445–487, 2005. 48

AL-KARAKI, J.; KAMAL a.E. Routing Techniques in Wireless Sensor Networks: A Survey. *IEEE Wireless Communications*, v. 11, n. 6, p. 6–28, 2004. ISSN 1536-1284. Disponível em: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1368893>. 55

ALCARAZ, C. et al. Wireless Sensor Networks and the Internet of Things : Do We Need a Complete Integration ? *1st International Workshop on the Security of the Internet of Things (SecIoT'10)*, n. June 2015, p. 1–8, 2010. 27, 42, 124

ANDRADE, A. et al. Outlier detection using k-means clustering and lightweight methods for Wireless Sensor Networks. In: *IECON Proceedings (Industrial Electronics Conference)*. [S.l.: s.n.], 2016. ISBN 9781509034741. 117, 174, 185

ATZORI, L.; IERA, A.; MORABITO, G. The Internet of Things: A survey. *Computer Networks*, Elsevier B.V., v. 54, n. 15, p. 2787–2805, 2010. ISSN 13891286. 41

BAHREPOUR, M. et al. Use of Event Detection Approaches for Outlier Detection in Wireless Sensor Networks. p. 439–444, 2009. 53, 71, 75, 150

BALL, G. H.; HALL, D. J. *ISODATA, a novel method of data analysis and pattern classification.* [S.l.], 1965. 82

BETTENCOURT, S. Separating the Wheat from the Chaff: Practical Anomaly Detection Schemes in Ecological Applications of Distributed Sensor Networks. In: *Distributed Computing in Sensor Systems.* [S.l.: s.n.], 2007. v. 4549, p. 223–239. ISBN 978-3-540-73089-7. 112, 114

BHOJANNAWAR, S. S.; BULLA, C. M.; DANAWADE, V. M. Anomaly Detection Techniques for Wireless Sensor Networks - A Survey. v. 2, n. 10, p. 3852–3857, 2013. 28, 34, 86, 92, 112, 133

BHOLOWALIA, P.; KUMAR, A. EBK-Means : A Clustering Technique based on Elbow Method and K-Means in WSN. *International Journal of Computer Applications*, v. 105, n. 9, p. 17–24, 2014. ISSN 09758887. Disponível em: <http://research.ijcaonline.org/volume105/number9/pxc3899674.pdf>. 116

BOANO, C. A. et al. The impact of temperature on outdoor industrial sensornet applications. *IEEE Transactions on Industrial Informatics*, v. 6, n. 3, p. 451–459, 2010. ISSN 15513203. 117

CALLEGARO, R. F. UMA ARQUITETURA PARA FUSÃO DE DADOS. 2014. 15, 64, 66, 100

CHANDOLA, V.; BANERJEE, A.; KUMAR, V. Anomaly detection. *ACM Computing Surveys*, v. 41, n. 3, p. 1–58, 2009. ISSN 03600300. 70, 94, 98, 121

CHEN, S.; ALMEIDA, L.; WANG, Z. A dynamic dual-rate beacon scheduling method of ZigBee/IEEE 802.15.4 for target tracking. In: *Proceedings - 2010 6th International Conference on Mobile Ad-hoc and Sensor Networks, MSN 2010.* [S.l.: s.n.], 2010. p. 103–109. ISBN 9780769543154. 117

CHENG, L. et al. A Survey of Localization in Wireless Sensor Network. *International Journal of Distributed Sensor Networks*, v. 2012, p. 1–12, 2012. ISSN 1550-1329. 28, 106

CHENG, L. et al. *A survey of localization in wireless sensor network*. 2012. 33

DASARATHY, B. V. Sensor fusion potential exploitation-innovative architectures and illustrative applications. *Proceedings of the IEEE*, v. 85, n. 1, p. 24–38, 1997. ISSN 00189219. 15, 59, 63, 64, 67

DING, X.; TIAN, Y.; YU, Y. A Real-Time Big Data Gathering Algorithm Based on Indoor Wireless Sensor Networks for Risk Analysis of Industrial Operations. *IEEE Transactions on Industrial Informatics*, v. 12, n. 3, p. 1232–1242, 2016. ISSN 15513203. 116

DUNN, J. C. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, 1973. ISSN 00220280. 82

DURRANT-WHYTE, H. F. *Sensor Models and Multisensor Integration*. 1988. 97–113 p. 15, 60, 64, 66

ELMENREICH, W. Sensor Fusion in Time-Triggered Systems. n. 9226605, 2002. 58, 60, 61, 133

ELMENREICH, W. Fusion of Continuous-valued Sensor Measurements using Confidence-weighted Averaging. *Journal of Vibration and Control*, v. 13, p. 1303–1312, 2007. ISSN 1077-5463. 102, 103, 143

EPOSS. *Internet of Things in 2020: A roadmap for the future*. [S.l.: s.n.], 2008. ISBN 3402823669209. 42

ESTRIN, D. et al. Next Century Challenges: Scalable Coordination in Sensor Networks. *Proceedings of the ACM/IEEE International Conference on Mobile Computing and Networking*, n. Section 4, p. 263–270, 1999. 46

FAROOQI, A. H.; KHAN, F. A. A survey of Intrusion Detection Systems for Wireless Sensor Networks. *International Journal of Ad Hoc and Ubiquitous Computing*, v. 9, n. 2, p. 69, 2012. ISSN 1743-8225. 44

FELSKE, M. S. et al. GLHOVE: A framework for uniform coverage monitoring using cluster-tree wireless sensor networks. In: *IEEE International Conference on Emerging Technologies and Factory Automation, ETFA*. [S.l.: s.n.], 2013. ISBN 9781479908622. ISSN 19460740. 115

FORGY, E. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics*, 1965. 82

GENTLE, J. E.; KAUFMAN, L.; ROUSSEUW, P. J. Finding Groups in Data: An Introduction to Cluster Analysis. *Biometrics*, 2006. ISSN 0006341X. 83

HALL, D. L. D. L.; MEMBER, S.; LLINAS, J. An introduction to multisensor data fusion. *Proceedings of the IEEE*, v. 85, n. 1, p. 6–23, 1997. ISSN 00189219. 28, 58

HE, Q. P.; WANG, J. Fault detection using the k-nearest neighbor rule for semiconductor manufacturing processes. *IEEE Transactions on Semiconductor Manufacturing*, v. 20, n. 4, p. 345–354, 2007. ISSN 08946507. 117

HEINZELMAN, W. B.; CHANDRAKASAN, A. P.; BALAKRISHNAN, H. An application-specific protocol architecture for wireless microsensor networks. *IEEE Transactions on Wireless Communications*, 2002. ISSN 15361276. 56

HILL, J. et al. System architecture directions for networked sensors. *ACM SIGOPS Operating Systems Review*, v. 34, n. 5, p. 93–104, 2000. ISSN 01635980. 45, 61, 84

HODGE, V. J.; AUSTIN, J. A Survey of Outlier Detection Methodoligies. *Artificial Intelligence Review*, v. 22, n. 1969, p. 85–126, 2004. 32, 34, 43, 70, 71, 98, 142

HONG, Z.; WANG, R.; LI, X. A clustering-tree topology control based on the energy forecast for heterogeneous wireless sensor networks. *IEEE/CAA Journal of Automatica Sinica*, v. 3, n. 1, p. 68–77, 2016. ISSN 23299274. 116

HUTCHISON, D.; MITCHELL, J. C. *Lecture Notes in Computer Science*. [S.l.: s.n.], 2004. ISSN 0302-9743. ISBN 9783642215148. 88

JAIN, A. K. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, v. 31, n. 8, p. 651–666, 2010. ISSN 01678655. 15, 79, 80, 81, 83, 117

JIN, R. et al. An rssi-based localization algorithm for outliers suppression in wireless sensor networks. *Wireless Networks*, v. 21, n. 8, p. 2561–2569, Nov 2015. ISSN 1572-8196. Disponível em: <https://doi.org/10.1007/s11276-015-0936-x>. 33, 127

JURDAK, R. et al. Wireless Sensor Network Anomalies: Diagnosis and Detection Strategies. *Intelligent Systems Reference Library*, v. 10, p. 309–325, 2011. ISSN 18684394. 28, 71

KEPHART, J. O.; CHESS, D. M. The vision of autonomic computing. *Computer*, v. 36, n. 1, p. 41–50, Jan 2003. ISSN 0018-9162. 122

KHATIRI, A.; MIRJALILY, G.; KHADEMZADEH, A. Energy-efficient shortcut tree routing in ZigBee networks. In: *Proceedings - 2012 4th International Conference on Computational Intelligence, Communication Systems and Networks, CICSyN 2012*. [S.l.: s.n.], 2012. p. 117–122. ISBN 9780769548210. 116

KIM, H.-S.; BANG, J.-S.; LEE, Y.-H. Distributed network configuration in large-scale low power wireless networks. *Computer Networks*, v. 70, p. 288–301, 2014. ISSN 13891286. Disponível em: <http://linkinghub.elsevier.com/retrieve/pii/S1389128614002217>. 28, 30, 116

KRISHNAMACHARI, B.; IYENGAR, S. Distributed Bayesian Algorithms for Fault-Tolerant Event Region Detection in Wireless Sensor Networks. p. 95, 2003. 88, 105, 150

LEÃO, E. et al. Superframe duration allocation schemes to improve the throughput of cluster-tree wireless sensor networks.

Sensors (Switzerland), v. 17, n. 2, 2017. ISSN 14248220. 50, 115, 171, 194

LEAO, E. et al. An allocation scheme for IEEE 802.15.4-ZigBee cluster-tree networks. In: *IECON Proceedings (Industrial Electronics Conference)*. [S.l.: s.n.], 2016. p. 4639–4644. ISBN 9781509034741. 51, 115

LI, C. et al. A Survey on routing protocols for large-scale wireless sensor networks. *Sensors*, v. 11, n. 4, p. 3498–3526, 2011. ISSN 14248220 (ISSN). Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-79953846552{&}partnerID=40{&}md5=07d2d6043da607b391b617bc0f>. 16, 48, 53, 54, 119

LINDSEY, S.; RAGHAVENDRA, C.; SIVALINGAM, K. Data gathering algorithms in sensor networks using energy metrics. *IEEE Transactions on Parallel and Distributed Systems*, 2002. ISSN 1045-9219. 57

LINDSEY, S.; RAGHAVENDRA, C. S. PEGASIS: Power-efficient gathering in sensor information systems. In: *IEEE Aerospace Conference Proceedings*. [S.l.: s.n.], 2002. ISBN 078037231X. ISSN 1095323X. 56

LOUREIRO, A. Grandes Desafios da Pesquisa em Computação no Brasil – 2006 – 2016. 2006. 53

LOUREIRO, A. A. F. et al. Redes de Sensores Sem Fio. *Simpósio Brasileiro de Redes de Computadores (SBRC)*, p. 179–226, 2003. Disponível em: <http://homepages.dcc.ufmg.br/{~}loureiro/cm/docs/sbrc03.p>. 15, 54

LUO, R. C. et al. Multisensor fusion and integration: approaches, applications, and future research directions. *Sensors Journal, IEEE*, v. 2, n. 2, p. 107–119, 2002. ISSN 1530-437X. 58, 66

MANJESHWAR, A.; AGRAWAL, D. P. TEEN: A Routing Protocol for Enhanced Efficiency in Wireless Sensor Networks.

Proceedings 15th International Parallel and Distributed Processing Symposium, 2001. ISSN 15302075. 57

MARKOS, M.; SINGH, S. Novelty detection: A review Part 1: Statistical approaches. *Signal Processing*, v. 83, p. 2481–2497, 2003. ISSN 01651684. 98

MARKOU, M.; SINGH, S. Novelty detection: A review - Part 2:: Neural network based approaches. *Signal Processing*, v. 83, p. 2499–2521, 2003. ISSN 01651684. 98

MARZULLO, K. Tolerating failures of continuous-valued sensors. v. 8, n. 4, p. 284–304, 1990. ISSN 07342071. 102, 133, 143

MERATNIA, N.; HAVINGA, P. Outlier Detection Techniques for Wireless Sensor Networks: A Survey. *IEEE Communications Surveys {&} Tutorials*, v. 12, n. 2, p. 159–170, 2010. ISSN 1553-877X. 34

MESMOUDI, A.; FEHAM1, M.; LABRAOUI, N. Wireless Sensor Networks Localization Algorithms: a Comprehensive Survey. *International Journal of Computer Networks & Communications*, v. 5, 6, n. 6, p. 45–64, 2013. 28

MOSHTAGHI, M. et al. Streaming analysis in wireless sensor networks. *Wireless Communications and Mobile Computing*, v. 14, p. n/a—-n/a, 2012. ISSN 15308669. 95

MOUSAVI, A. et al. Spatio-temporal event detection using probabilistic graphical models (PGMs). *2013 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, p. 81–88, 2013. 110, 114

NAKAMURA, E. F.; LOUREIRO, A. A. F.; FRERY, A. C. Information Fusion for Wireless Sensor Networks: Methods, Models, and Classifications. *ACM Computing Surveys*, v. 39, n. 3, p. 1–55, 2007. ISSN 03600300. Disponível em: <http://search.ebscohost.com/login.aspx?direct=true{&}db=bth{&}AN=26561958{&}site=eds-live{&}aut>. 51, 60, 62, 150

OLADIMEJI, M. O.; SMIEE, M. G.; MIEEE, S. D. A New Approach for Event Detection using k -means Clustering and Neural Networks. p. 1–5, 2015. 16, 111, 114

PAN, M. S.; TSENG, Y. C. Quick convergecast in ZigBee beacon-enabled tree-based wireless sensor networks. *Computer Communications*, v. 31, n. 5, p. 999–1011, 2008. ISSN 01403664. 48

PANWAR, A.; KUMAR, S. A. Localization Schemes in Wireless Sensor Networks. *2012 Second International Conference on Advanced Computing & Communication Technologies*, p. 443–449, 2012. Disponível em: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6168410>. 33, 48

PEI, X. et al. Spatio-temporal Event Detection: A Hierarchy Based Approach for Wireless Sensor Network. *2014 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery*, p. 372–379, 2014. 15, 75, 109, 112, 150

PELLEG, D.; MOORE, A. X-means: Extending K-means with Efficient Estimation of the Number of Clusters. *CEUR Workshop Proceedings*, 2015. ISSN 16130073. 83

PINTO, A. R. et al. An approach to implement data fusion techniques in wireless sensor networks using genetic machine learning algorithms. *Information Fusion*, v. 15, n. 1, p. 90–101, 2014. ISSN 15662535. 117

RASSAM, M. A.; ZAINAL, A.; MAAROF, M. A. *Advancements of data anomaly detection research in Wireless Sensor Networks: A survey and open issues*. 2013. 30, 32, 44, 46, 84, 86, 92, 95, 133, 142, 143, 150

RASSAM, M. a.; ZAINAL, A.; MAAROF, M. A. An adaptive and efficient dimension reduction model for multivariate wireless sensor networks applications. *Applied Soft Computing*, Elsevier B.V., v. 13, n. 4, p. 1978–1996, 2013. ISSN 15684946. 94, 112

ROSS, S. M.; PH, D. Peirce ' s criterion for the elimination of suspect experimental data. *Journal of Engineering Technology*, v. 20, p. 1–12, 2003. ISSN 07479964. 15, 99, 100, 133, 143

RUSSELL, S.; NORVIG, P. *Artificial Intelligence: A Modern Approach, 3rd edition*. [S.l.: s.n.], 2009. ISSN 0269-8889. ISBN 0136042597. 77

SAMPAIO, R.; MANCINI, M. Estudos de revisão sistemática : um guia para síntese. *Revista Brasileira de Fisioterapia*, v. 11, p. 83–89, 2007. ISSN 1413-3555. 213

SCHÖLKOPF, B.; SMOLA, A.; MÜLLER, K. R. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Computation*, 1998. ISSN 08997667. 83

SHU, L.; MUKHERJEE, M.; WU, X. Toxic gas boundary area detection in large-scale petrochemical plants with industrial wireless sensor networks. *IEEE Communications Magazine*, v. 54, n. 10, p. 22–28, 2016. ISSN 01636804. 117

SINGH, S. P.; SHARMA, S. Range Free Localization Techniques in Wireless Sensor Networks: A Review. *Procedia Computer Science*, v. 57, p. 7–16, 2015. ISSN 18770509. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1877050915018864>. 33

TAYLOR, J. R. *Introdução à Análise de Erros*. [S.l.: s.n.], 2012. 352 p. ISBN 9788540701373. 101, 133, 143

VELMANI, R.; KAARTHICK, B. An efficient cluster-tree based data collection scheme for large mobile wireless sensor networks. *IEEE Sensors Journal*, v. 15, n. 4, p. 2377–2390, 2015. ISSN 1530437X. 28

XIE, R.; JIA, X. Transmission-efficient clustering method for wireless sensor networks using compressive sensing. *IEEE Transactions on Parallel and Distributed Systems*, v. 25, n. 3, p. 806–815, 2014. ISSN 10459219. 116

YAO, Y. et al. A RSSI-based distributed weighted search localization algorithm for WSNs. *International Journal of Distributed Sensor Networks*, 2015. ISSN 15501477. 33, 127

YETGIN, H. et al. *A Survey of Network Lifetime Maximization Techniques in Wireless Sensor Networks*. 2017. 27

YICK, J.; MUKHERJEE, B.; GHOSAL, D. Wireless sensor network survey. *Computer Networks*, v. 52, n. 12, p. 2292–2330, 2008. ISSN 13891286. 15, 45, 46

YIN, J.; HU, D.; YANG, Q. Spatio-Temporal Event Detection Using Dynamic Conditional Random Fields. *Ijcai*, n. 3, p. 1321–1326, 2009. 75, 150

YOUSSEF, A. M.; YOUSSEF, M. A Taxonomy of Localization Schemes for Wireless Sensor Networks. *Icwn*, v. 9, n. 8, p. 1754–1757, 2007. ISSN 18125638. 106

ŽALIK, K. R. An efficient k-means clustering algorithm. *Pattern Recognition Letters*, v. 29, n. 9, p. 1385–1391, 2008. ISSN 01678655. 79, 82, 132, 152

ZENG, Y. et al. Secure localization and location verification in wireless sensor networks: A survey. *Journal of Supercomputing*, v. 64, n. 3, p. 685–701, 2013. ISSN 09208542. 33

ZHANG, Y. et al. Statistics-based outlier detection for wireless sensor networks. *International Journal of Geographical Information Science*, v. 26, n. May, p. 1373–1392, 2012. ISSN 1365-8816. 30, 32

ZHANG, Y. Z. Y.; MERATNIA, N.; HAVINGA, P. Outlier Detection Techniques for Wireless Sensor Networks: A Survey. *IEEE Communications Surveys {&} Tutorials*, v. 12, n. 2, 2010. ISSN 1553-877X. 29, 43, 44, 73, 84, 86, 92, 98, 99, 104, 112, 131, 133, 142

ZHOU, B.; CAO, Z.; GERLA, M. Cluster-based inter-domain routing (CIDR) protocol for MANETs. In: *WONS 2009 - 6th International Conference on Wireless On-demand Network Systems and Services*. [S.l.: s.n.], 2009. ISBN 9781424433759. 57

ZHU, C. et al. A tree-cluster-based data-gathering algorithm for industrial WSNs with a mobile sink. *IEEE Access*, v. 3, p. 381–396, 2015. ISSN 21693536. 115

Zigbee Alliance. *ZigBee/IEEE 802.15.4 specification*. [S.l.: s.n.], 2011. v. 2011. 314 p. ISBN 9780738166834. 15, 46, 49, 51

Apêndices

APÊNDICE A – Revisão Sistematizada

A revisão bibliográfica seguiu um modelo de estudo sistemático (SAMPAIO; MANCINI, 2007), com o objetivo de reunir estudos semelhantes publicados sobre o tema de maneira imparcial e completa. As etapas seguidas foram descritas em termos pesquisas realizadas. A revisão é dividida em introdução, metodologia, resultados e conclusão.

A.1 Introdução

Dentre questionamentos desta tese, o principal é de como melhorar o funcionamento de uma RSSF de larga escala. Para tanto consideramos os principais elementos de influência como: formação topológica da redes, dados livres de anomalias (qualidade do dados), volume de tráfego na rede, periodicidade do monitoramento e consumo energético dos nodos.

Com estes elementos em vista, alguns questionamentos da pesquisa surgem, os quais já foram apresentados no Capítulo 1, mas, por conveniência, são reproduzidos de forma sucinta aqui:

- É possível melhorar a detecção, identificação e tratamento de *outliers* em RSSF de larga escala considerando um ambiente sujeito a interferências e falhas?

- O uso de técnicas para agrupamento de dados pode contribuir para melhorar o conhecimento de áreas monitoradas, separando-as e permitindo focar nas mais prioritárias (sujeitas à ocorrência de eventos relevantes)?
- Dado que em redes de larga escala usualmente os nodos não conseguem alcançar o coordenador em um único salto, como a formação de rotas baseadas em dados pode contribuir para a reduzir o atraso fim-a-fim, o consumo energético e o volume de dados de RSSF de larga escala?
- Dado que uma RSSF de larga escala produz grandes volumes de dados, como as técnicas de *Big data* podem contribuir para melhorar o desempenho da rede?

Esses questionamentos visam nortear esta revisão sistemática, limitando a e sua abrangência na busca de artigos e trabalhos acadêmicos.

A.2 Metodologia

As pesquisas foram realizada nas base de dados do *Google Scholar* e *IEEE Xplore*, sendo a primeira a principal fonte e a segunda complementar. Para as buscas foram utilizadas os questionamentos como ponto inicial, limitador e com vistas ao foco de pesquisa do Programa de Pós-graduação de Engenharia de Automação e Sistemas da UFSC.

Dessa forma, limitou-se a pesquisa nas seguintes questões: retirada de anomalias dos dados monitorados, clusterização de dados e formação de *cluster-tree* baseado em dados.

Alguns critérios de pesquisa adicionais foram usados, devido a grande número de resultados nos artigos relacionados, visando reduzir o total de artigos. Este processo foi necessário dado que o assunto pesquisado é comum a outras áreas e, portanto, é útil para selecionar os trabalhos mais relevantes.

Foram utilizados três operadores nas pesquisas, "+" para indicar que os resultados devem conter o termo associado, "-" para remover qualquer artigo que contenha o termo associado e as (" ") para indicar que o termo associado deve ser idêntico ao descrito no interior das aspas, incluindo espaço e símbolos.

As principais buscas são listadas nas Seção A.3, na qual em cada linha subsequente são adicionadas os termos listados são indicados o número de artigos encontrados à direita e utiliza-se a letra k para indicação de milhar. As ramificações finais de cada busca são indicadas em cores diferentes (negrito) e utilizam todos os termos anteriores aos membros.

Ao final de cada ramificação é utilizado um protocolo de inclusão/exclusão, no qual, primeiramente foram lidos os títulos e resumos selecionando os artigos relevantes à área de busca. Complementarmente, foram incluídos os artigos com assuntos relacionados diretamente ao trabalho, os artigos que possuíam apresentação de resultados com critérios para seleção de parâmetros de simulação, os artigos que possuíam critérios de avaliação e também de áreas semelhantes que poderiam fornecer *insight* sobre o trabalho.

A.3 Resultados

Os resultados da pesquisa foram divididos em duas partes para abranger maior quantidade de termos que envolvam os elementos desta arquitetura.

Nas Tabelas 25 e 26 são listados os trabalhos sobre detecção de outlier em RSSF com topologia *cluster-tree.*

Tabela 25 – Terminologia e número de artigos encontrados na busca utilizando o *Google Scholar*.

Termos	Resultados
WSN + "outlier detection"	1.7k
WSN + "outlier detection" + "cluster-tree"	36
Após leitura dos Resumos	**32**

Tabela 26 – Terminologia e número de artigos encontrados na busca utilizando o IEEE Xplore.

Termos	Resultados
WSN + "outlier detection"	318
WSN + "outlier detection" + "Cluster-tree"	5
Após leitura dos Resumos	**5**

Nas Tabelas 27 e 28 são listados os trabalhos sobre clusterização de dados em RSSF com topologia cluster-tree.

Após a etapa inicial da busca, foram selecionados 69 artigos mencionados. Na sequência, foram consultados referências utilizadas nestes artigos e pesquisados trabalhos relacionados dos mesmos autores, somando mais 12 artigos e totalizando 81 artigos.

Neste revisão sistemática foram selecionadas as

Tabela 27 – Terminologia e número de artigos encontrados na primeira busca utilizando o *Google Scholar*.

Termos	**Resultados**
WSN + "data clustering"	3.5K
WSN + data clustering + Cluster-tree	53
Após leitura dos Resumos	**27**

Tabela 28 – Terminologia e número de artigos encontrados na primeira busca utilizando o IEEE *xplore*.

Termos	**Resultados**
WSN + "data clustering"	193
WSN + "data clustering" + Cluster-tree	6
Após leitura dos Resumos	**5**

principais ocorrências para o tema desta tese. Para tal, foi dividido a pesquisa em duas frentes para aumentar a abrangência do tema. Os principais artigos e seus resumos críticos foram evidenciados nos trabalhos relacionados na Seção 3.6.

No sentido da detecção de outlier e clusterização de dados para RSSF, os trabalhos indicam a necessidade de uma melhor integração entre os temas. Uma vez que estes temas individualmente possuem um vasto referencial bibliográfico, no entanto sua junção pode ser melhor explorar alcançando melhores resultados.

Por outro lado os artigos indicam uma lacuna científica a ser explorada, apesar dos trabalhos no sentido da formação de topologias *cluster-tree*: baseado no consumo energético, no balanceamento do volume de dados e dos

híbridos. Ainda assim, fica evidente a carência de pesquisa na área da formação de *cluster-tree* baseados em dados que possibilitem novas abordagens formação de redes *cluster-tree*.