MARIA MERCEDES GAMBOA-MEDINA

**DETECÇÃO DE VAZAMENTOS EM REDES SOB PRESSÃO
BASEADA NA ANÁLISE DOS SINAIS DE VAZÃO E PRESSÃO COM UM SISTEMA
DE RECONHECIMENTO DE PADRÕES**

Dissertação apresentada à Escola de Engenharia
de São Carlos, da Universidade de São Paulo,
como parte dos requisitos para obtenção do
título Mestre em Ciências: Engenharia
Hidráulica e Saneamento.

Orientadora: Prof. Tit. Luisa Fernanda Ribeiro
Reis

VERSÃO CORRIGIDA

SÃO CARLOS, SP

2013

Este trabalho foi licenciado com uma Licença Creative Commons - Atribuição – Não Comercial 3.0 Brasil.

É permitido compartilhar e adaptar o trabalho para todo uso não comercial, desde que respeitada a atribuição. A licença jurídica (resumida ou completa) pode ser acessada em: <http://creativecommons.org/licenses/by-nc/3.0/br/deed.pt>

Gamboa-Medina, Maria Mercedes

G192d Detecção de vazamentos em redes sob pressão baseada na análise dos sinais de pressão e vazão com um sistema de reconhecimento de padrões / Maria Mercedes Gamboa-Medina; orientadora Luisa Fernanda Ribeiro Reis. São Carlos, 2013.

Dissertação (Mestrado) - Programa de Pós-Graduação e Área de Concentração em Hidráulica e Saneamento -- Escola de Engenharia de São Carlos da Universidade de São Paulo, 2013.

1. Vazamentos. 2. Redes. 3. Sinais de pressão. 4. Análise de sinais. I. Título.

FOLHA DE JULGAMENTO

Candidata: Engenheira **MARIA MERCEDES GAMBOA MEDINA**.

Título da dissertação: "Detecção de vazamentos em redes sob pressão baseada na análise dos sinais de vazão e pressão com um sistema de reconhecimento de padrões".

Data da defesa: 29/07/2013

Comissão Julgadora:

Resultado:

Profa. Titular **Luisa Fernanda Ribeiro Reis (Orientadora)**
(Escola de Engenharia de São Carlos/EESC)

Aprovado

Prof. Dr. **Rodrigo Capobianco Guido**
(Universidade Estadual Paulista "Júlio de Mesquita Filho"/UNESP –
São José do Rio Preto)

Aprovado

Prof. Dr. **Alexandre Kepler Soares**
(Universidade Federal de Goiás/UFG)

Aprovado

Coordenadora do Programa de Pós-Graduação em Engenharia Hidráulica e
Saneamento:

Profa. Assoc. **Maria Bernadete A. Varesche Silva**

Presidente da Comissão de Pós-Graduação:
Prof. Titular **Denis Vinicius Coury**

Ao Javier, o homem maravilhoso que decidiu comigo andar a vida em parceria, e a cada dia faz que eu queira criar muitos caminhos novos com ele.

À minha mãe Blanca, meu pai Rafael, meus irmãos Rafael, Juan Francisco (i.m.), Ricardo e Alejandro, e minha irmã Laura, que estiveram e estarão sempre presentes na minha vida e que, para meu orgulho, fazem parte da minha própria identidade; e aos meus sobrinhos, que carregam tanto futuro e esperança.

AGRADECIMENTOS

Agradeço sinceramente às muitas pessoas que de diversas maneiras contribuíram para o grande processo que me permite apresentar esta dissertação. Em particular gostaria de expressar minha gratidão pela contribuição durante o projeto:

À professora Fernanda, que desde o início acreditou neste projeto e na minha capacidade, e me fez acreditar também. Agradeço pela permanente orientação e o conhecimento transmitido, e também pela confiança depositada, a disposição e paciência dispensadas, e os ensinamentos proporcionados com suas palavras e com seu exemplo.

Ao professor Rodrigo Guido, cujas indicações foram uma bússola nas diversas etapas do projeto, agradeço a grande contribuição e a excelente disposição em colaborar conosco.

Aos professores Rodrigo Porto, André de Carvalho e Maria Carolina Monard, agradeço pela receptividade e sua contribuição resolvendo dúvidas e oferecendo indicações que foram fundamentais no trabalho.

Ao engenheiro Javier Mahecha, pelas ideias fornecidas, a valiosa colaboração nas atividades de laboratório, as longas conversas sobre os temas do projeto, a revisão do texto e, em fim, tudo o suporte que fez possível a conclusão deste trabalho.

A Roberto Cezar Bérghamo e Fransico Bergamin, técnicos do departamento de hidráulica e saneamento, pela colaboração com as reparações no experimento mostrando sua competência profissional e excelente disposição.

Ao Casiano, funcionário da firma TeS, pelo esforçado trabalho na reforma do sistema eletrônico do experimento e a colaboração frente as dificuldades correlatas.

Aos colegas do Laboratório de Simulação Numérica, particularmente a Narumi Abe pela ajuda nas atividades de laboratório e na implementação das análises, a Frederico Odan pela disposição para resolver dúvidas e a Tais Shinma pela paciência e amizade.

Aos colegas do ICMC, Didier Vega e Lilian Berton, pelo auxílio no uso do *software*.

Aos colegas, funcionários e professores do departamento de hidráulica e saneamento, pelo acolhimento, a companhia, o auxílio e os ensinamentos proporcionados durante o mestrado.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pela bolsa de estudos concedida.

Acredito nos homens, isto é, na sua razão. Sem essa fé não teria a força de sair da cama pela manhã. É, eu acredito na calma força da razão nos homens. Não poderão resistir a ela muito tempo, não poderão aguentar. Pensar é um dos maiores prazeres da raça humana.

A vida de Galileu. Bertol Bretsch (1858 – 1956)

RESUMO

GAMBOA-MEDINA, M. M. **Deteção de vazamentos em redes sob pressão baseada na análise dos sinais de pressão e vazão com um sistema de reconhecimento de padrões.**

Dissertação (Mestrado) - Programa de Pós-Graduação e Área de Concentração em Hidráulica e Saneamento -- Escola de Engenharia de São Carlos da Universidade de São Paulo, 2013.

O controle de perdas em sistemas de distribuição de água para abastecimento é uma preocupação constante, e uma tarefa fundamental para a solução do problema é a detecção rápida e confiável dos vazamentos que frequentemente iniciam em qualquer ponto da rede. Uma abordagem promissória é a detecção de vazamentos baseada na análise de sinais adquiridos pelo monitoramento das redes durante sua operação, e dentro dela se enquadra este trabalho. É desenvolvido um sistema de reconhecimento de padrões para análise de sinais de pressão e vazão que permite identificar se durante a aquisição do sinal aconteceu um vazamento ou não. Para a conformação desse sistema diversas técnicas são exploradas, incluindo a extração de características no domínio do tempo (energia, entropia, número de cruzamentos por zero) e na decomposição *wavelet* (distribuição da energia nas componentes). Também é explorado o uso de algoritmos para classificação de diferentes tipos (vizinhos mais próximos, árvore de decisão, regra de decisão, Naive Bayes, máquina de vetor suporte e rede neural artificial com funções de base radial). Sinais são adquiridos junto ao circuito hidráulico experimental, que permitiu a simulação da ocorrência de um vazamento na rede, para constituir uma ampla base de dados com sinais de exemplo. Além da revisão bibliográfica e os conceitos relativos às metodologias exploradas, são apresentadas neste documento as análises que conduzem à criação do sistema de reconhecimento de padrões mais apropriado para o problema. Das análises dos diferentes métodos considerados é definido o sistema de reconhecimento de padrões, em suas etapas de segmentação e padronização, extração de características e classificação. A avaliação do sistema proposto mostra um desempenho totalmente satisfatório, com reconhecimento acertado de sinais vinculados ou não a um vazamento em mais de 95% dos testes.

Palavras chave: Vazamentos. Redes de abastecimento de água. Sinais de pressão. Análise de sinais.

ABSTRACT

GAMBOA-MEDINA, M. M. Leak detection in water networks based on the analysis of flow and pressure signals by a pattern recognition system. Master thesis – Escola de Engenharia de São Carlos of Universidade de São Paulo, 2013.

Control of losses in water supply systems is a constant concern, and a key to the solution of this problem is the rapid and reliable detection of leaks that often begin anywhere on the network. A promising approach to solve the problem is the leak detection based on the analysis of signals acquired by monitoring the network in operation, and this research fits with that approach. It's developed a pattern recognition system for the analysis of pressure and flow signals, which identifies whether a leak happened during signal acquisition. Several techniques are exploited for forming this system, including the feature extraction in the time domain (energy, entropy, zero crossings count) and in the wavelet decomposition (energy distribution in the components). Also, the use of different types of algorithms for classification (nearest neighbors, decision tree, decision rule, Naive Bayes, support vector machine and artificial neural network with radial basis functions) is explored. Signals are acquired from the experimental hydraulic circuit, allowing the simulation of the onset of a leak in the network, to form a big database of example signals. Besides the literature review and the concepts relating to the considered methods, in this document are shown the analyses leading to the creation of the pattern recognition system most appropriate for the problem. The analysis of the methods considered allows defining the pattern recognition system, which is composed by segmentation and standardization, feature extraction and classification. The evaluation of the proposed system shows a completely satisfactory performance, recognizing rightly the signals as linked or not to a leak in more than 95% of the tests.

Keywords: Leaks. Supply networks. Pressure signals. Signal analysis.

LISTA DE ABREVIATURAS

AAC	Área abaixo da curva ROC, indicador de desempenho.
AS_O	Métodos baseados na análise de sinais durante operação.
AS_T	Métodos baseados na análise de sinais durante o transitório.
COM	Nome da classe que agrupa os sinais adquiridos durante o início de um vazamento.
db2	Função <i>wavelet</i> Daubechies com dois momentos zero.
EDW	Distribuição da energia nas componentes da decomposição <i>wavelet</i> , característica do sinal.
ENE	Energia, característica do sinal.
ENT	Entropia, característica do sinal.
FBR	Rede neural artificial com funções de base radial, método de classificação.
FC	Falsos COM, número de instâncias da classe SEM classificadas na classe COM.
FS	Falsos SEM, número de instâncias da classe COM classificadas na classe SEM.
KNN	K vizinhos mais próximos, método de classificação.
MH_P	Métodos baseados na modelagem hidráulica no regime permanente.
MH_T	Métodos baseados na modelagem hidráulica no regime transitório.
NCZ	Número de cruzamentos por zero, característica do sinal.
RNA	Rede neural artificial.
ROC	<i>Receiver operating characteristic</i> , curva que representa o desempenho.
SEM	Nome da classe que agrupa os sinais adquiridos durante escoamento permanente.
SVM	Máquina de vetor suporte (<i>support vector machine</i>), método de classificação.
TOD	Vetor combinando as características ENE, ENT, NCZ e EDW.
VC	Verdadeiros COM, número de instâncias classificadas acertadamente na classe COM.
VS	Verdadeiros SEM, número de instâncias classificadas acertadamente na classe SEM.

LISTA DE FIGURAS

Figura 1. Esquema geral dos elementos do modelo hidráulico	24
Figura 2. Esquema geral dos métodos baseados na análise de sinais.....	29
Figura 3. Função <i>wavelet</i> Daubechies, db2	41
Figura 4. Esquema da decomposição <i>wavelet</i> discreta.....	43
Figura 5. Esquema da decomposição <i>wavelet packet</i>	44
Figura 6. Esquema do circuito hidráulico experimental.....	47
Figura 7. Localização dos sensores no circuito hidráulico experimental	49
Figura 8. Interface do <i>software</i> de aquisição de dados.....	50
Figura 9. Três exemplos de sinais da classe SEM adquiridos com o sensor P02 para três diferentes vazões no circuito (Q).....	51
Figura 10. Três exemplos de sinais da classe COM adquiridos com o sensor P02 para três diferentes vazões no circuito (Q) e diferentes magnitudes de vazamento (V).	52
Figura 11. Esquema básico das etapas de um sistema de reconhecimento de padrões	54
Figura 12 Exemplo de dois sinais de pressão adquiridos.	58
Figura 13 Características extraídas dos dois sinais de pressão.....	58
Figura 14. Espaço de entrada para elementos de dimensão três pertencentes a duas classes ..	62
Figura 15. Exemplo de árvore de decisão para duas classes, com quatro características.	64
Figura 16. Esquema de rede de Funções de Base Radial	70
Figura 17. Exemplo curva ROC	73
Figura 18. Esquema das análises apresentadas neste capítulo, em relação às componentes do sistema de reconhecimento de padrões.....	76
Figura 19. AAC média usando diferentes níveis de decomposição, segundo o classificador..	78
Figura 20. Comparação da AAC média usando decomposição <i>wavelet</i> discreta e <i>wavelet packet</i>	80
Figura 21. Histograma de frequências da distribuição da energia na decomposição <i>wavelet packet</i> , para exemplos SEM e COM.	81
Figura 22. AAC usando variações do algoritmo K-NN, depois da extração das características TOD.....	84
Figura 23. AAC usando variações do algoritmo K-NN, depois da extração da característica EDW.	85
Figura 24. Histogramas dos valores do vetor de características para exemplos SEM e COM.	88

Figura 25. AAC para diferentes vetores de características segundo o sensor e o classificador considerados.....	91
Figura 26. AAC para diferentes classificadores, segundo o sensor.	97
Figura 27. Curvas ROC para os sensores P02, P16 e Q03.....	103
Figura 28. Sensibilidade (taxa de verdadeiros positivos) para diferentes valores de V/Q, total.	107

LISTA DE TABELAS

Tabela 1. Resumo de trabalhos em detecção de vazamentos	23
Tabela 2. Relações V/Q dos exemplos COM.....	52
Tabela 3. Esquema da matriz de confusão	71
Tabela 4. AAC média (desvio padrão) para diferentes níveis de decomposição, segundo o classificador.	78
Tabela 5. Aumento percentual na AAC com o aumento no nível de decomposição, segundo o classificador.	79
Tabela 6. AAC média (desvio padrão) usando decomposição <i>wavelet</i> packet, segundo o classificador.	80
Tabela 7. AAC usando variações do algoritmo K-NN, depois da extração do vetor de características TOD.	84
Tabela 8. AAC usando variações do algoritmo K-NN, depois da extração da característica EDW.	85
Tabela 9. Incremento percentual na AAC usado variações do algoritmo KNN.....	86
Tabela 10. Intervalos de valores mais frequentes das características.....	87
Tabela 11. AAC média (desvio padrão) para diferentes vetores de características, segundo o sensor considerado.....	92
Tabela 12. AAC média para diferentes vetores de características, segundo o classificador.	93
Tabela 13. AAC média (desvio padrão) segundo o sensor considerado, para os vetores de características de melhor desempenho.	94
Tabela 14. AAC média (desvio padrão) para diferentes classificadores, segundo o sensor considerado.....	96
Tabela 15. Matrizes de confusão totais para cada sensor, e média.	100
Tabela 16. Taxa de acerto, taxa de erro e sensibilidade do sistema, para cada sensor.....	101
Tabela 17. Taxa de falsos positivos, especificidade e precisão do sistema, para cada sensor e média.	102
Tabela 18. AAC do sistema para cada sensor, e média.	104
Tabela 19. Indicadores VC e FS do sistema de reconhecimento calculados para os diferentes valores de V/Q usados nos testes.....	106
Tabela 20. Sensibilidade (taxa de verdadeiros positivos) para os diferentes valores de V/Q usados nos testes.....	107

SUMÁRIO

1. INTRODUÇÃO	17
2. OBJETIVOS	21
3. MÉTODOS DE DETECÇÃO DE VAZAMENTOS	22
3.1. MÉTODOS BASEADOS EM MODELOS HIDRÁULICOS (MH)	24
3.1.1. Modelos em regime permanente (MH_P)	25
3.1.2. Modelos em regime transitório (MH-T)	27
3.1.3. Vantagens e limitações dos métodos MH	28
3.2. MÉTODOS DE DETECÇÃO DE VAZAMENTOS BASEADO NA ANÁLISE DE SINAIS (AS)	29
3.2.1. Análise de sinais de pressão durante um transitório hidráulico (AS_T)	29
3.2.2. Análise dos sinais durante operação da rede (AS_O)	33
4. CONCEITOS DE ANÁLISE DE SINAIS E APRENDIZADO DE MÁQUINA	38
4.1. VARIÁVEIS DO SISTEMA COMO SINAIS	38
4.2. TRANSFORMAÇÕES DE DOMÍNIO	39
4.2.1. Domínio da frequência e Transformada de Fourier	40
4.2.2. Transformada <i>Wavelet</i> e análise tempo-frequência	41
4.3. DECOMPOSIÇÃO <i>WAVELET</i> DISCRETA E <i>PACKET</i>	43
4.4. APRENDIZADO DE MÁQUINA	45
5. AQUISIÇÃO DE SINAIS EXPERIMENTAIS EM LABORATÓRIO	46
5.1. MONTAGEM EXPERIMENTAL	46
5.2. AQUISIÇÃO DE DADOS	48
5.3. RESUMO DOS EXPERIMENTOS	50

6.	SISTEMA DE RECONHECIMENTO DE PADRÕES.....	53
6.1.	SEGMENTAÇÃO E PADRONIZAÇÃO	54
6.2.	EXTRAÇÃO DE CARACTERÍSTICAS	57
6.2.1.	Energia (ENE).....	58
6.2.2.	Entropia (ENT)	59
6.2.3.	Número de cruzamentos por zero (NCZ).....	59
6.2.4.	Distribuição da energia na decomposição <i>wavelet</i> (EDW).....	60
6.3.	CLASSIFICADOR	61
6.3.1.	Vizinhos mais próximos	61
6.3.2.	Árvore de decisão	63
6.3.3.	Regras de Decisão.....	65
6.3.4.	Naive Bayes	66
6.3.5.	Máquina de vetores de suporte	67
6.3.6.	Rede Neural Artificial de Funções de Base Radial.....	68
6.4.	AVALIAÇÃO DO PROCESSO.....	70
6.4.1.	Validação cruzada (<i>k-fold cross validation</i>)	70
6.4.2.	Matriz de confusão.....	71
6.4.3.	Curva ROC	73
6.5.	IMPLEMENTAÇÃO DO SISTEMA DE RECONHECIMENTO DE PADRÕES..	74
7.	RESULTADOS	76
7.1.	SELECÇÃO DO TIPO E NÍVEL DE DECOMPOSIÇÃO <i>WAVELET</i> PARA A CARACTERÍSTICA EDW	77
7.2.	PARÂMETROS PARA O CLASSIFICADOR KNN	83
7.3.	CARACTERÍSTICAS EXTRAÍDAS	87
7.4.	TIPOS DE CLASSIFICADOR.....	95
7.5.	RESULTADOS FINAIS DO SISTEMA DE RECONHECIMENTO DE PADRÕES	98
7.5.1.	Matriz de confusão.....	99
7.5.2.	Área abaixo da curva ROC (AAC)	102

7.6.	MAGNITUDE DE VAZAMENTO	104
8.	CONCLUSÕES	108
9.	REFERÊNCIAS	112
APÊNDICE A.	HISTOGRAMAS DOS VALORES DAS CARACTERÍSTICAS EXTRAÍDAS DOS SINAIS PARA TODOS OS SENSORES	116
APÊNDICE B.	DESEMPENHO DO SISTEMA DE RECONHECIMENTO DE PADRÕES PROPOSTO, RESULTADOS INDIVIDUAIS DA VALIDAÇÃO CRUZADA	125
APÊNDICE C.	CURVA ROC DO CLASSIFICADOR NO SISTEMA DE RECONHECIMENTO DE PADRÕES PROPOSTO (TODOS OS SENSORES)	129

1. INTRODUÇÃO

Em todo o mundo, o gerenciamento de recursos hídricos para atender as demandas de abastecimento urbano está sendo cobrado a aumentar sua eficiência, e um ponto chave no atendimento desse propósito é o controle dos vazamentos. A incidência de vazamentos em redes de abastecimento de água no Brasil, mesmo que não seja quantificada com certeza, é reconhecidamente elevada. Segundo dados do Sistema Nacional de Informações sobre Saneamento (SNSA, 2010) as perdas totais (reais e aparentes) são em média de 37,4%, atingindo 53,4% quando avaliada por regiões do país e até 80,7% ao considerar as empresas prestadoras de serviços. As consequências dos vazamentos vão além da evidente perda econômica pela água captada, tratada e transportada pelo sistema mas não aproveitada nem faturada. Os vazamentos também estão relacionados com perdas de energia (especialmente em sistemas com bombeamento), desestabilização do solo com prejuízo das obras de infraestrutura e edificações, e problemas de saúde pública pelo risco de intrusão de contaminantes na rede, além de constituir um desperdício do recurso, contrário à racionalidade e sustentabilidade que o sistema deve objetivar. Apesar do grande interesse suscitado em relação aos vazamentos, eles representam um problema de difícil solução em todos os sistemas de abastecimento de água ao redor do mundo, de forma que é impossível dizer que na atualidade existe alguma cidade com sua rede de distribuição em condições tão controladas que não apresente vazamento. Uma das razões para isso é o fato de que existem níveis de controle de vazamentos economicamente inviáveis (ASHTON; HOPE, 2001).

As indesejadas consequências dos vazamentos localizados, alvos dos estudos do presente trabalho, aumentam segundo o volume de água perdido até que sejam tomadas providências para a sua correção, isto é, o tempo transcorrido que inclui a detecção de um vazamento, sua localização e reparação. O tempo de reparação depende da capacidade logística, técnica e econômica do prestador do serviço, e comumente sua redução encontra menores dificuldades técnicas do que a detecção e localização. Assim, estas duas tarefas iniciais são as mais influentes nos sistemas razoavelmente organizados, e constituem alvos das metodologias mencionadas neste trabalho.

A prática antiga de esperar que manifestações evidentes de um vazamento sejam detectadas pelos usuários ou inspetores para só depois tomar providências no sentido de resolver o problema, persiste em um número de sistemas urbanos cada vez mais reduzido. Os exames

visuais têm sido substituídos em muitos casos por outros mais avançados como a inspeção acústica, injeção de gás na rede, inspeções eletromagnéticas ou baseadas em medições elétricas. A incorporação das novas tecnologias tem ampliado as possibilidades de detectar vazamentos menores que os identificáveis através da simples inspeção visual, mas continuam tendo um limite abaixo do qual os vazamentos não podem ser detectados.

Considerando que os vazamentos que podem ser detectados pelos usuários ou pelas consequências facilmente perceptíveis no sistema, podem corresponder a menos de 10% das perdas reais em sistemas bem manejados (PILCHER, 2007), as tarefas de detecção e localização são complexas e de grande importância. Os vazamentos não visíveis, tanto os detectáveis via metodologias de inspeção quanto os de magnitudes menores à detectável através delas, estão caracterizados por vazões pequenas, ausência de indicadores visíveis de superfície e influência menos evidente nas características hidráulicas do sistema. Assim, os tempos de reparo desses vazamentos são bem maiores do que os dos visíveis, e, portanto, ainda que caracterizados por vazões pequenas, podem promover maior volume de água perdida.

As metodologias baseadas em procedimentos diretos sobre a rede, como as anteriormente mencionadas, são consideradas metodologias passivas. Dentre as suas limitações estão a dificuldade de obtenção de informação atualizada de todo o sistema, o fato dos resultados dependerem da experiência dos operadores e das condições de realização das inspeções, e, em alguns casos, a necessidade da suspensão do serviço de abastecimento aos usuários. Além disso, os dados para análise geralmente estão limitados a um período curto (durante a realização de inspeções ou testes) e possuem pequena abrangência espacial. Face ao exposto, a sua utilidade prática se restringe à localização de vazamentos previamente detectados.

Assim, é cada vez mais importante desenvolver o controle ativo de vazamentos, entendido como o processo de detecção e localização de vazamentos permanente sobre a rede (WLTS, 2007). Para que novas metodologias de controle ativo possam ser implementadas elas devem ser condizentes com a realidade dos órgãos responsáveis pelo abastecimento, o que inclui informação incompleta sobre as características e estado da rede, pouca disponibilidade de pessoal altamente qualificado, recursos limitados para a compra de equipamentos sofisticados e fragilidade das redes existentes, dentre outros fatores.

Na procura de um controle ativo de vazamentos, diversas metodologias têm sido propostas no âmbito da engenharia hidráulica. A premissa fundamental é de que o vazamento é um fenômeno hidráulico e as suas consequências têm interferência no funcionamento do sistema,

e em razão disso podem ser analisadas, por exemplo, em termos de características como vazão e pressão. Esta abordagem tem sido explorada por muitos pesquisadores nas últimas décadas, com propostas diversas, mas a sua aplicação prática a sistemas reais tem sido alcançada em poucos casos.

Por outro lado, os esforços pela modernização dos sistemas de distribuição de água para abastecimento visando a melhorar seu controle e operação têm levado à instalação de sensores de baixo custo que permitem conhecer o estado de sistema, por exemplo, vazões e pressões em locais específicos. Essas informações são obtidas continuamente, mas, em geral, são pobremente analisadas.

Considerando a necessidade do controle ativo de vazamentos e a possibilidade de aquisição de informações, existem propostas baseadas na análise do comportamento das variáveis do sistema, isto é, na análise dos sinais adquiridos. Este enfoque mostra-se especialmente útil por permitir que vazamentos sejam detectados assim que surgirem, e por não depender da modelagem hidráulica, o que implica numa vantagem importante considerando as dificuldades práticas para a utilização de modelos de redes de distribuição reais.

O presente trabalho é um aporte dentro dos esforços para a detecção de vazamentos pela análise de sinais. Usando os dados de pressão e vazão adquiridos experimentalmente num circuito de laboratório, é desenvolvido um sistema de reconhecimento de padrões para a classificação dos sinais adquiridos durante uma situação de vazamento ou não. O embasamento teórico, a explicação da metodologia utilizada e os resultados obtidos são apresentados neste documento, que é organizado nos seguintes capítulos:

- Objetivos, ao longo do qual são especificados os objetivos do trabalho;
- No capítulo 3 é apresentada uma revisão da bibliografia sobre métodos de detecção de vazamentos no âmbito da engenharia hidráulica, segundo uma classificação aqui proposta das técnicas baseadas nos modelos hidráulicos e das baseadas na análise dos sinais. São comentados os trabalhos mais representativos dos diferentes enfoques, e o tipo de sistemas para os quais as técnicas têm sido validadas;
- No capítulo 4 são apresentados brevemente, para o leitor que não está familiarizado com a terminologia da área de análise de sinais ou de aprendizado de máquina, os conceitos básicos que serão aplicados nos capítulos seguintes.

- No capítulo 5 são descritas as condições sob as quais foi adquirida a informação utilizada nas análises, incluindo o sistema de laboratório e aquisição e os experimentos executados.
- Os conceitos relativos ao sistema de reconhecimento de padrões com suas três etapas principais (segmentação e padronização, extração de características e classificação) são apresentados no capítulo 6. Igualmente, são descritas as alternativas consideradas dentro de cada etapa do sistema de reconhecimento.
- O capítulo 7 contém os resultados e as discussões correspondentes. São apresentadas as análises para a escolha das metodologias internas no sistema de reconhecimento de padrões, segundo as alternativas definidas no capítulo 6, assim como os resultados finais obtidos com a arquitetura de sistema selecionada e ainda análises sobre a influência da magnitude do vazamento na detecção.
- As conclusões e recomendações são apresentadas no capítulo respectivo.
- Ao final, as referências bibliográficas empregadas são listadas.
- Nos apêndices, dados e figuras adicionais aos colocados nos diferentes capítulos são mostrados.

2. OBJETIVOS

O objetivo deste trabalho é o desenvolvimento de uma metodologia para detecção de vazamentos em redes de água sob pressão baseada na análise dos sinais de pressão e vazão em pontos pré-escolhidos da rede, com a construção de um sistema de reconhecimento de padrões. É utilizado um circuito experimental e adquiridos em laboratório os sinais posteriormente utilizados para a construção e teste do sistema.

Os objetivos específicos são:

- Adaptação do circuito hidráulico experimental para estudo das características do escoamento: vazão em trechos da rede e pressão em pontos escolhidos;
- Obtenção de sinais de pressão e vazão característicos do sistema em diferentes estados de funcionamento, na ausência e presença de vazamentos, executando os experimentos de interesse sobre o circuito;
- Identificação da arquitetura do sistema de reconhecimento de padrões para a detecção de vazamentos, considerando os diferentes métodos alternativos que podem ser empregados nas suas três etapas: segmentação e padronização, extração de características e classificação.
- Implementação da rotina computacional para cada uma das três etapas do sistema, incluindo os métodos alternativos para cada qual.
- Avaliação das alternativas para a etapa de extração de características, analisando e comparando quatro diferentes características extraídas do sinal, três delas no domínio do tempo e uma depois da aplicação da transformada *wavelet*.
- Avaliação das alternativas para a etapa de classificação, analisando seis métodos ou classificadores.
- Seleção da melhor arquitetura do sistema de reconhecimento de padrões e avaliação completa do seu desempenho.
- Avaliação da influência da magnitude do vazamento no desempenho do sistema.

3. MÉTODOS DE DETECÇÃO DE VAZAMENTOS

Em resposta ao desafio da detecção e localização de vazamentos, diversas abordagens têm sido propostas. Por exemplo, ainda são comumente usadas as do tipo mais antigo: as inspeções, que evoluíram desde os exames visuais para o uso de sensores acústicos, injeção de gás, medições eletromagnéticas e uma diversidade de técnicas mais ou menos sofisticadas. As metodologias baseadas em procedimentos diretos sobre a rede como as anteriores são consideradas metodologias passivas. Dentre as suas limitações está o fato de que é muito difícil obter informação atualizada de todo o sistema, os resultados dependem da experiência dos executores e das condições de execução, as informações se restringem ao período de execução e possuem pequena abrangência espacial, e em alguns casos implicam na interrupção do serviço aos usuários. Face ao exposto, a sua utilidade prática se restringe à localização de vazamentos previamente detectados.

Na procura de um controle ativo de vazamentos, outras metodologias têm sido desenvolvidas. Os vazamentos têm sido reconhecidos como fenômenos hidráulicos cuja ocorrência influi no estado do sistema hidráulico onde acontecem, e conseqüentemente têm impacto sobre variáveis como a vazão e pressão. O estudo dessa influência visando à detecção indireta dos vazamentos tem dado origem a vários grupos de técnicas propostas no âmbito da engenharia hidráulica. A seguir são apresentadas brevemente essas técnicas, sem pretender abarcar a totalidade dos trabalhos da área e sim alguns avaliados como representativos. Um resumo desses trabalhos com suas principais características, e organizados segundo a classificação proposta, é apresentado na Tabela 1.

Para compreender o panorama atual das técnicas de detecção de vazamentos é proposta neste trabalho uma nova classificação, distinguindo as técnicas baseadas na modelagem hidráulica (MH) das técnicas focadas na análise dos sinais (AS). Dentro do primeiro grupo estão as técnicas que modelam o sistema em regime permanente (MH_P) ou transitório (MH_T). Dentro do segundo grupo podem se diferenciar as técnicas AS_T que estudam o sistema somente durante um transitório hidráulico (situações cabíveis a modelos dinâmicos inerciais) e as técnicas AS_O que analisam os sinais durante a operação normal de uma rede, numa escala temporal muito superior à da passagem das ondas e em condições de análise quase-estáticas. É dentro deste último subgrupo que poderia ser classificado o sistema de reconhecimento de padrões desenvolvido no presente trabalho.

Tabela 1. Resumo de trabalhos em detecção de vazamentos

	Autor, Ano	Método Principal	Sistema considerado	Avaliação		
				Num	Lab	Cam
MH	Métodos Baseados em Modelos Hidráulicos					
MH_P	Modelos em regime Permanente					
	Pudar; Liggett, 1992	Problema Inverso	Circuitos hipotéticos	X	-	-
	Andersen; Powell, 2000	Estimacion de estado. Minimos cuadrados	Circuitos hipotéticos	X		
	Soares, 2003	Problema Inverso. AGs	Circuitos hipotéticos	X	-	-
	Poulakis; Valougeorgis; Papadimitriou, 2003	Probabilidad Bayesiana e modelos parametricos	Circuitos hipotéticos	X	-	-
	Colombo, 2007	Problema Inverso. AGs	Setor de rede real	-	-	X
	Wu, 2009	Problema Inverso. AGs	Setor de rede real	-	-	X
	Wu; Sage; Turtle, 2010	Problema Inverso. AGs	Circuitos hipotéticos e Setor de rede real	X	-	X
MH_T	Modelos Em Regime Transitório					
	Liggett; Chen, 1994	Análise transitório inverso (ITA)	Circuitos hipotéticos	X		
	Vitkovsky; Simpson; Lambert, 2000	Análise transitório inverso (ITA) e AGs	Circuitos hipotéticos	X		
	Kapelan; Savic; Walters, 2003	Análise transitório inverso (ITA), AGs e LVM	Circuitos hipotéticos	X	-	-
	Saldarriaga; Fuentes; Galvis, 2006	Análise transitório inverso (ITA) e AGs	Setor de rede real	-	-	X
	Soares, 2007	Análise transitório inverso (ITA) e AGs	Circuitos hipotéticos e de laboratório	X	X	-
	Covas; Ramos, 2010	Análise transitório inverso (ITA)	Sistema aberto de laboratório	-	X	-
	Soares; Covas; Reis, 2011	Análise transitório inverso (ITA), AGs e LVM	Circuito laboratório	-	X	-
AS	Métodos Baseados Na Analise De Sinais					
AS_T	Análise De Sinais De Pressão Durante Um Transitório Hidráulico					
	Mpesha; Gassman; Chaudhry, 2001	Fluxo oscilatorio. Resposta em frecuencia	Sistema aberto hipotético	X	-	-
	Wang et al., 2002	Amortecimento da onda. Harmonicos	Sistema tubulação única hipotético e de laboratório	X	X	-
	Ferrante; Brunone, 2003	Teste transitorío. Análise harmônico	Sistema tubulação única hipotético e de laboratório	X	X	-
	Ferrante; Brunone, 2003	Teste transitorío. Análise wavelet	Sistema tubulação única de laboratório	-	X	-
	Misiunas, 2003; Misiunas et al., 2005	Sumas cumulativas (CUSUM)	Sistema tubulação única de laboratório e real	-	X	X
	Covas; Ramos; Almeida, 2005	Fluxo oscilatorio. Resposta em frecuencia	Sistema aberto hipotético	X	-	-
	Ferrante; Brunone; Meniconi, 2007	Teste transitorío. Análise wavelet	Sistema tubulação única hipotético e de laboratório	X	X	-
	Ferrante; Brunone; Meniconi, 2009	Teste transitorío. Análise wavelet	Sistema aberto (junção) de laboratório e real	-	X	X
	Aguiar, 2010	Classificação com RNA	Sistema tubulação única de laboratório	-	X	-
Meniconi et al., 2011	Teste transitorío. Análise wavelet e Lagrange	Sistema aberto (junção) real	-	-	X	
AS_O	Análise De Sinais Durante Operação Da Rede					
	Buchberger; Nadimpalli, 2004	Analise estatistico de vazões	Setor de rede hipotético e real	X	-	X
	Aksela; Aksela; Vahala, 2009	Mapas auto organizados	Setor de rede real	-	-	X
	Romano; Kapelan; Savic, 2009	Controle estatistico e inferencia Bayesiana	Setor de rede real	-	-	X
	Mounce; Boxall; Machell, 2010	RNA e Logica Difusa	Setor de rede real	-	-	X
	Ye; Ph; Fenner, 2011	Filtros Kalman	Setor de rede real	-	-	X
	Palau; Arregui; Carlos, 2012	Analise de componentes principais	Setor de rede real	-	-	X
	Romano; Kapelan; Savic, 2012	RNA, Controle estatistico e inferencia Bavesiana	Setor de rede real	-	-	X

3.1. MÉTODOS BASEADOS EM MODELOS HIDRÁULICOS (MH)

Considerando os vazamentos como fenômenos hidráulicos, diversos autores têm explorado a sua incorporação nos modelos hidráulicos. O modelo hidráulico é a representação, simplificada em maior ou menor grau, de um sistema composto de elementos (condutos, válvulas, bombas, etc.) que definem o escoamento de água no interior das redes sob pressão. É assumido que as demandas e as características da rede (tais como geometria, material e comprimento das tubulações, cotas, fator de atrito, etc) determinam o estado do sistema, em termos de vazão em cada trecho e pressão em cada nó, sob dada condição operacional. Como esquematizado na Figura 1, as relações entre as características da rede, demandas e estado do sistema são preestabelecidas através das equações da hidráulica que descrevem a conservação de massa nos nós e de energia nos circuitos fechados.

Figura 1. Esquema geral dos elementos do modelo hidráulico

Assim, a modelagem hidráulica pode ser vista como problema direto, quando todas as características da rede, demandas e condições operacionais são conhecidas, enquanto as pressões e vazões são incógnitas. Ocorre que algumas características da rede ou sofrem variações ao longo do tempo, como é o caso das rugosidades absolutas dos condutos, ou não são conhecidas por razões diversas. Assim, uma abordagem muito usada em pesquisas recentes é a do problema inverso (PUDAR; LIGGETT, 1992), em que valores de pressão e vazão, normalmente obtidos por medição na rede, são fornecidos ao modelo para determinar algumas características ou demandas desconhecidas. A resolução do problema inverso tem sido aplicada, por exemplo, para calibrar o fator de atrito, ou pesquisar vazamentos, assumidos como demandas (exemplos indicados na Tabela 1 serão comentados nas seções seguintes).

Os métodos para a resolução do modelo também podem ser classificados em dois tipos, explícitos ou implícitos. Nos métodos explícitos, o funcionamento hidráulico do sistema é descrito por N equações não lineares, onde estão envolvidas as características da rede e as demandas, e por N incógnitas representando o estado do sistema. Os valores das incógnitas são encontrados através da resolução do sistema de equações. Poucos problemas podem ser

representados fielmente com esta abordagem simplificada, mas tem sido utilizada para a calibração do fator de atrito e as demandas nos nós usando desde métodos de tentativa e erro até resoluções complexas do sistema de equações da hidráulica da rede (SOARES, 2003).

Nos métodos implícitos, o problema inverso é resolvido usando o equacionamento, as características conhecidas da rede e os valores de pressão e vazão obtidos. São procurados os valores ótimos das características desconhecidas com o objetivo de minimizar o erro, que é definido como a diferença entre a resposta do sistema (pressão, vazão) calculada pelo modelo e os seus valores observados (LIGGETT; CHEN, 1994).

A modelagem hidráulica pode ser feita para o regime permanente (MH_P) ou não permanente (MH_T), sendo este o critério de classificação das técnicas de detecção de vazamentos referidas a seguir.

3.1.1. Modelos em regime permanente (MH_P)

Nos modelos hidráulicos de redes em regime permanente, é suposto que as mudanças na rede são instantâneas, passando de um estado permanente para outro, desconsiderando a transição. As equações que descrevem o funcionamento hidráulico da rede sob regime permanente são conhecidas e aceitas para propósitos diversos tais como projetos, reabilitação, e otimização da operação de sistemas de distribuição de águas para abastecimento. Existe atualmente uma grande quantidade de simuladores hidráulicos no regime permanente, com diferentes graus de complexidade e limitações, alguns deles desenvolvidos por grupos de pesquisa das universidades (*Crede*), outros por agências públicas (*Epanet*), ou companhias comerciais (*WaterCad*, *InfoWater*, *H2OMap*). No regime permanente é factível estudar simultaneamente pressão e vazão, considerando a capacidade de sensores e sistemas de aquisição, diferentemente das análises em regime transitório, como discutido na seção 3.1.2.

A detecção de vazamentos por MH_P pode ser abordada com a incorporação de demandas ou saídas adicionais e desconhecidas nos nós da rede. Os valores de tais demandas ou saídas, que podem ser maiores ou iguais a zero, podem ser também variáveis de decisão do problema de otimização utilizado para calibração de redes (quando usada a abordagem implícita para resolução do problema inverso (PUDAR; LIGGETT, 1992)). Assim, é em algumas pesquisas sobre calibração de modelos que podem ser encontrados os primeiros esforços para a detecção de vazamentos.

Diferentes trabalhos sobre a calibração de modelos incluindo a determinação de demandas foram desenvolvidos nas últimas décadas, muitos deles identificando os algoritmos genéticos (AGs) como uma ferramenta vantajosa aos propósitos de otimização, mas somente alguns incluíram dentro do problema de calibração as perdas por vazamentos e demandas dependentes da pressão (COLOMBO, 2007; SOARES, 2003).

Por exemplo, na proposta apresentada por Wu (2009) e Wu, Sage e Turtle (2010), os vazamentos dependentes da pressão são incluídos no modelo hidráulico e a definição dos parâmetros, em cada nó, é essencialmente um problema de calibração, resolvido com AGs. Nesse último trabalho é apresentada a aplicação do método a uma rede teórica e a um setor de rede real com dados adquiridos no período de mínima demanda obtendo, segundo os autores, resultados positivos, mesmo limitados pelas dificuldades da modelagem hidráulica e da acurácia na medição. O modelo de vazamentos incorporado ao modelo hidráulico define uma relação potencial da vazão de saída do vazamento com a pressão. Usualmente essa relação pode ser dada pela equação de um orifício ou por formulações específicas que mudam segundo a magnitude e tipo do vazamento.

Mas nem todos os trabalhos envolvendo modelação hidráulica de redes de abastecimento de água com inclusão de vazamentos para escoamento permanente estão ligados à calibração; essa abordagem tem sido desenvolvida também para aplicações diferentes, como a otimização da operação e/ou localização de válvulas redutoras de pressão (REIS; PORTO; CHAUDHRY, 1997). Focando no problema da detecção de vazamentos, Poulakis, Valougeorgis e Papadimitriou (2003) apresentam uma técnica na qual não é considerada a calibração, mas sim as grandes incertezas envolvidas na simulação de redes reais. A ideia geral é calcular as saídas de um modelo hidráulico do sistema dependente das diferentes condições prováveis de vazamento, para posteriormente compará-las com um conjunto de observações de características do sistema (vazão, pressão) e escolher a condição mais provável. As condições de vazamento avaliadas, que são a posição e a magnitude dos vazamentos, são as variáveis de decisão do problema de otimização, enquanto a probabilidade dessas variáveis corresponderem à saída observada é a função objetivo a maximizar mediante um processo de busca. Análises probabilísticos bayesianos dão suporte ao método, que foi testado para uma rede teórica.

Com uma abordagem menos comum, Anderson e Powell (2000) apresentaram uma metodologia para detecção de vazamentos baseada na estimação de estado, isto é, o cálculo de pressões e vazões, dados os parâmetros fixos do sistema, usando método implícito de

mínimos quadrados ponderados. São utilizadas as equações do fluxo em circuitos e as variáveis de estado são as demandas (vazamentos) desconhecidas nos nós.

3.1.2. Modelos em regime transitório (MH-T)

O estudo do comportamento hidráulico no regime não permanente tem alcançado maior relevância nos últimos anos, motivado em parte pela necessidade de representar fenômenos comuns nas redes como variabilidade das demandas e operação de válvulas, bombas e reservatórios. Uma vantagem importante das análises em regime transitório é o fornecimento de maior quantidade de informações sobre o sistema, dado que em cada ponto de análise conta-se com grande quantidade de dados, variáveis no tempo. Consequentemente, o problema inverso torna-se sempre indeterminado (PUUST et al., 2010). Assim, a análise hidráulica de sistemas durante o transitório hidráulico tem sido aplicada a diversas metodologias para a calibração de modelos e detecção de vazamentos.

O princípio básico utilizado é que, quando um transitório acontece, uma onda se movimenta pelo sistema com umas características de propagação dadas, até encontrar uma descontinuidade (como um vazamento, mudança de diâmetro, mudança de material, etc.); por efeito de dita descontinuidade as características da onda são modificadas, e essa modificação é perceptível também em outros pontos do sistema, pela propagação. Assim, estudando o sinal original e a modificação que sofreu, é possível identificar a singularidade que a produz e sua localização.

A ideia central antes exposta é aplicada no método de análise do transitório inverso (ITA, do inglês *Inverse Transient Analysis*) (LIGGETT; CHEN, 1994), procurando a grandeza e localização de vazamentos no modelo hidráulico de transitório, que minimizem a diferença entre o diagrama de pressão em função do tempo modelado e o medido no sistema. Autores posteriores têm feito aportes a esta abordagem, usando como metodologias de otimização os algoritmos genéticos (VITKOVSKY; SIMPSON; LAMBERT, 2000), o método de Levenberg-Marquardt (LVM), ou métodos híbridos (KAPELAN; SAVIC; WALTERS, 2003) (SOARES; COVAS; REIS, 2011).

Mesmo sendo uma técnica bem estudada, são poucos os relatos na literatura sobre sua aplicação prática, como mostrado na Tabela 1. Uma mesma rede teórica simples foi considerada na apresentação de várias das metodologias propostas (KAPELAN; SAVIC; WALTERS, 2003; LIGGETT; CHEN, 1994; VITKOVSKY; SIMPSON; LAMBERT, 2000),

também existem casos de estudo baseados em experimentação em laboratório com circuitos fechados (SOARES; COVAS; REIS, 2011; SOARES, 2007) ou abertos (COVAS; RAMOS, 2010), enquanto são excepcionais as referências a testes em sistemas reais como o conseguido por Saldarriaga, Fuentes e Galvis (2006). A importância da precisão no modelo hidráulico é a principal limitação para a aplicação efetiva do método de transitório inverso a redes reais, e os seus resultados não podem ser apresentados sem uma quantificação da incerteza envolvida (PUUST et al., 2010). A construção de um modelo hidráulico requer um conhecimento detalhado e preciso do sistema físico a ser modelado, condição difícil de ser atingida por redes reais.

Adicionalmente, a modelagem de redes no regime transitório é ainda foco de pesquisa. Soares (2007) desenvolveu uma metodologia inovadora para calibração e detecção de vazamentos, em um modelo hidráulico em regime transitório incorporando o fator de atrito variável com o comportamento viscoelástico das tubulações de PVC, e os vazamentos como saídas governadas pela pressão.

Por outro lado, os métodos baseados no regime não permanente apresentam uma limitação adicional devido a que para implementá-los é necessário gerar fenômenos transitórios no sistema a avaliar, o que no caso de redes de abastecimento implicaria em grande risco de falha. Além disso, existe a dificuldade de estudar as vazões durante o transiente, dado que a maioria de medidores não responde instantaneamente às mudanças, limitando assim as análises aos registros de pressões (PUUST et al., 2010).

3.1.3. Vantagens e limitações dos métodos MH

Analisando todas as técnicas de detecção de vazamentos MH (baseadas na modelagem hidráulica do sistema, tanto no regime permanente quanto no não permanente) podem-se identificar algumas vantagens e limitações gerais. As vantagens principais são a possibilidade de obter informação detalhada sobre o fenômeno e suas consequências em toda a extensão da rede, assim como permitir a análise igualmente detalhada de cenários criados, o que significa uma ferramenta muito útil para o planejamento e a tomada de decisões. Dentre as limitações pode-se apontar a necessidade de uma série de condições difíceis de completar para uma rede de abastecimento real: alto detalhamento na modelação, alta quantidade e qualidade da informação sobre o sistema, exatidão nos parâmetros e condições de contorno, exatidão na simulação hidráulica, e sistema de medição confiável e sincronizado.

Na maioria das redes urbanas, durante a construção e as operações de manutenção, são decididas mudanças que não são registradas nos documentos oficiais. Muitas delas nem podem ser esclarecidas em estudos posteriores considerando a localização das tubulações sob a infraestrutura da cidade, de profundidade incerta. Assim sendo, é muito difícil conseguir um conhecimento preciso e completo das redes, o que gera sempre um nível de incerteza mesmo no melhor modelo desenvolvido.

3.2. MÉTODOS DE DETECÇÃO DE VAZAMENTOS BASEADO NA ANÁLISE DE SINAIS (AS)

Visando superar as limitações dos métodos de detecção de vazamentos MH, têm surgido diversos trabalhos que focam o estudo na análise das variáveis dependentes, isto é de vazões e pressões, que podem ser obtidas por medições no sistema e consideradas genericamente sinais (seção 4.1). As técnicas classificadas neste grupo certamente reconhecem os mesmos princípios hidráulicos aplicados nos modelos, mas não focam os esforços na resolução das suas equações e sim na extração de informações diretamente dos sinais, sem passar pela modelagem hidráulica, para a detecção de vazamentos. Na maioria dos casos, é realizado algum tipo de comparação entre os sinais monitorados em qualquer intervalo de tempo e aqueles correspondentes ao estado original da rede, ou seja, sem vazamentos. Mesmo baseados nessa ideia fundamental, as técnicas que analisam os sinais durante o transitório hidráulico (AS_T) e aquelas focadas na análise durante a operação (AS_O) têm raios de ação bem diferenciados.

Figura 2. Esquema geral dos métodos baseados na análise de sinais.

3.2.1. Análise de sinais de pressão durante um transitório hidráulico (AS_T)

Da mesma forma que as técnicas MH-T (seção 3.1.2), nas técnicas AS_T é aproveitada a mudança que um vazamento causa na propagação da onda durante um fenômeno transitório. Quando acontece um transitório hidráulico gerado por uma manobra no sistema, e, durante a

sua propagação a onda encontra um vazamento este terá um efeito que pode ser de amortecimento e reflexão (COLOMBO; LEE; KARNEY, 2009). Também o próprio surgimento de um vazamento gera um transitório hidráulico e ondas que se deslocam pelo sistema. Identificar e quantificar todas essas alterações e correlacioná-las com um vazamento é o foco dos métodos de detecção AS_T que, diferentemente dos métodos MH_T, utilizam as informações monitoradas em um único ponto do sistema ou cada um dos pontos monitorados é analisado isoladamente.

Entre as técnicas AS_T propostas nos últimos anos está a detecção de rupturas fazendo uso do algoritmo CUSUM (do inglês *Cumulative Sum*) na análise de sinais de pressão (MISIUNAS, 2003; MISIUNAS et al., 2005). Essa técnica visa identificar a onda de pressão produzida por uma ruptura que dê origem a um vazamento para extrair dela informações sobre a localização e magnitude do vazamento, relacionados com os tempos de chegada e as amplitudes das ondas. Para tanto, são usadas informações de monitoramento permanente da pressão em um único ponto de uma tubulação em circuito aberto.

Outro tipo de análise de sinais de pressão durante o transitório gerado pelo próprio vazamento foi aplicado por Aguiar (2010), utilizando redes neurais artificiais (RNA) para a identificação de padrões de mudança no sinal relacionados com o início de um vazamento. A técnica visa distinguir sinais de pressão como pertencendo a uma de duas classes: referente ao início de um vazamento ou sem vazamento. Usando sinais de pressão adquiridos durante a simulação da ocorrência de vazamentos com abertura de válvulas num sistema experimental, foi treinada uma rede neural do tipo *Perceptron feedforward*. A RNA tem como entrada um segmento do sinal, i.e. um conjunto de valores de pressão adquiridos (e normalizados), e como saída dois valores cada um deles relativo à semelhança com uma das classes possíveis. Esse processo pode ser visto como um aprendizado de máquina supervisionado.

A análise de sinais pode ser realizada sobre diferentes representações ou domínios, como explicado na seção 4.2. Os dois trabalhos antes referidos analisam os sinais na sua representação original, no domínio temporal, enquanto outros muitos autores têm proposto métodos aplicando transformações de domínio para, por exemplo, fazer as análises de resposta em frequência ou de harmônicos. A ideia anterior é aplicada nos trabalhos referidos a seguir, nos quais, adicionalmente, as informações relativas à presença de vazamentos são investigadas durante a ocorrência de fenômenos transitórios controlados e gerados artificialmente, diferentes dos produzidos pelo próprio início do vazamento.

Aplicando o método de resposta em frequência, Mpesha, Gassman e Chaudhry (2001) estudam o comportamento hidráulico de um sistema aberto submetido ao escoamento oscilatório gerado pela operação cíclica de uma válvula na saída. Segundo as características do sistema e da oscilação, é calculado o diagrama de resposta em frequência no ponto a montante da válvula, considerando as oscilações da pressão e da vazão, e usando para tanto o método de matrizes de transferência. Considerando que um vazamento reflete as ondas de pressão, criando ondas superpostas e, portanto, alterando a resposta em frequência, essa resposta alterada pela existência de um vazamento é também calculada com o mesmo método analítico. Ao comparar as respostas em termos de frequência para os dois casos, é possível identificar os picos adicionais gerados pelo vazamento, e partindo de tais informações é definida uma metodologia para calcular a localização e a magnitude do vazamento. Nesse trabalho são apresentados testes numéricos com diferentes sistemas abertos incluindo tubulações em série ou em paralelo.

A mesma metodologia de geração de fluxo oscilatório é utilizada no método de ondas estacionárias apresentado por Covas, Ramos e Almeida (2005) também baseado na análise da resposta em frequência do sistema, mas calculando-a com o método de impedância. A análise de ressonância é usada para identificar as frequências geradas por um ou vários vazamentos, e daí a sua localização na tubulação. A metodologia é apresentada com exemplos numéricos de sistemas abertos de tubulação única ou tubulações em série.

Wang et al. (2002) propõem o estudo do amortecimento da onda de pressão gerada por um transitório hidráulico como ferramenta para detectar, localizar e quantificar um vazamento. O amortecimento da onda devido à fricção é estudado em uma tubulação sem nenhum vazamento, seja mediante análise numérica ou medição direta das pressões num ponto, e posteriormente é estudado o mesmo sistema com presença de vazamento. Os sinais de pressão assim obtidos são separados em segmentos relacionados com o período da onda, a forma de separar as partes correspondentes a cada reflexão da onda, e a cada um desses segmentos é aplicada a transformada de Fourier. No trabalho citado é desenvolvida analiticamente uma expressão para o amortecimento do transiente, sendo exatamente exponencial se o amortecimento é devido unicamente ao atrito e aproximadamente exponencial se o vazamento também contribui para o amortecimento. A amplitude de cada um dos componentes harmônicos resultantes da transformada é analisada juntamente com os parâmetros de amortecimento, de forma de obter as conclusões buscadas. É feita a apresentação numérica do método e também uma validação em teste de laboratório com tubulação única.

Em um estudo de validação da técnica proposta por Wang et al. (2002) concluiu-se que sua aplicação está limitada a sistemas que não envolvem nenhuma descontinuidade interna e a fenômenos transitórios instantâneos e de pequena amplitude cuja duração seja menor do que o tempo de viagem da onda até o ponto de monitoramento (NIXON et al., 2006).

Uma técnica AS_T destacada que usa transitórios hidráulicos gerados por manobras de fechamento rápido e análise de harmônicos é apresentada por Ferrante e Brunone (2003a). Estudando no domínio da frequência as equações que descrevem o escoamento transitório, obtém-se a expressão analítica para as pressões na saída da tubulação. Esse sinal de saída no domínio da frequência depende das características da manobra que produz o transitório e das características do sistema, e é modificado também pela presença de um vazamento. Sinais de pressão são adquiridos sob o transitório gerado controladamente e são transformados ao domínio da frequência usando transformada de Fourier, para posteriormente compará-los com a resposta analítica ou com a de experimentos anteriores. A análise comparativa permite a identificação de distorções que indicam a presença de vazamento na tubulação. A metodologia foi aplicada a um sistema aberto de tubulação única, com válvula de fechamento na saída, e medição das pressões em alta frequência de aquisição.

Ao estudar a resposta do sistema no domínio da frequência, é perdida uma porção de informações interessantes como, por exemplo, o tempo de chegada da onda ou descontinuidades no sinal temporal. Visando superar essa perda, num trabalho complementar (FERRANTE; BRUNONE, 2003b), foi proposta a alternativa de transformação das pressões de saída usando transformada *wavelet* contínua. No sinal transformado são identificadas as singularidades devidas à reflexão da onda pelo vazamento que permitem a sua detecção, e também é obtida informação sobre o tempo de chegada da onda refletida, indicativa da localização do vazamento. Os resultados da aplicação do método e análises realizadas sobre um sistema experimental aberto empregando diferentes tipos de transformada *wavelet* permitem concluir que nem todos têm a mesma utilidade para o objetivo de identificação das singularidades relativas ao vazamento.

Continuando a pesquisa do trabalho anterior, têm sido publicados estudos sobre o uso de *wavelets* para a análise de sinais de pressão em transitórios com o objetivo da detecção de vazamentos (FERRANTE; BRUNONE; MENICONI, 2007, 2009). Considerando um sistema com uma única tubulação, são gerados analiticamente sinais de pressão em resposta a um transitório hidráulico, usando para isso o método das características. Aos sinais é aplicada a transformada *wavelet* discreta e contínua, e vários tipos de *wavelets* mãe (esses conceitos são

apresentados na seção 4.2.2). São apresentados testes com séries geradas analiticamente e obtidas em laboratório e, embora tenha sido identificada a necessidade de investigação dos tipos de *wavelet* mais adequados à análise, é provada a capacidade da transformada *wavelet* para a identificação de singularidades a partir de sinais de pressão, mesmo na presença de ruído.

A análise anterior pode ser estendida de sistemas de uma única tubulação a sistemas com bifurcação (FERRANTE; BRUNONE; MENICONI, 2009), e ainda melhorias são propostas pela integração de modelos de *Lagrange*. O uso desses modelos Lagrangianos para avaliar as causas das singulares detectadas com a análise *wavelet* (MENICONI et al., 2011) foi testado em campo em uma tubulação adutora.

A totalidade das técnicas AS_T encontradas considera unicamente sistemas abertos, na maioria dos casos consistindo de uma única tubulação, ocasionalmente sistemas com uma bifurcação, como mostrado na Tabela 1. Tais técnicas, além de serem dependentes da análise de um evento transitório bem controlado e conhecido, são altamente sensíveis a irregularidades ou mudanças no sistema. Adicionalmente, existem dificuldades para aplicar as considerações analíticas em sistemas mais complexos. Por outro lado, existe também uma limitação importante pelo risco ao impor altas pressões durante os transitórios gerados em sistemas reais.

Assim, sem desconhecer os avanços no seu desenvolvimento e as vantagens apresentadas, pode-se concluir que as técnicas AS_T não são aplicáveis para verificação permanente da rede de abastecimento visando à detecção de vazamentos, sendo limitadas a linhas adutoras e à inspeção detalhada de trechos específicos.

3.2.2. Análise dos sinais durante operação da rede (AS_O)

Pesquisas recentes estão tentando aplicar os aprendizados dos métodos anteriormente resumidos para superar suas limitações, visando ao aproveitamento do monitoramento permanente da rede para a detecção de vazamentos pela análise dos sinais adquiridos durante a operação (AS_O). Este tipo de método procura ser aplicável a sistemas reais de distribuição, aproveitando as informações de monitoramento permanente e de baixo custo, sem requerer operações específicas na rede (como a geração de transitórios, que podem ainda prejudicar estruturalmente o sistema) ou o conhecimento detalhado necessário para a modelagem hidráulica de elevada precisão. Comumente são aplicadas ferramentas estatísticas avançadas

ou técnicas de inteligência artificial, e em geral são procuradas nos sinais continuamente monitorados evidências que revelem o início de um vazamento, e não a sua pré-existência como em muitas das técnicas AS_T.

Podem ser citados vários trabalhos que analisam os sinais de vazões de entrada em um setor de rede de abastecimento, como os resumidos a seguir. Análises estatísticas da vazão no período de baixa demanda (considerada como variável aleatória) são a base da proposta de Buchberger e Nadimpalli (2004), que é validada somente para setores residenciais pequenos, onde possam ser descartados consumos noturnos importantes e variáveis. Segundo o método, a partir dos dados de vazão de entrada a uma área residencial adquiridos durante o período de mínima demanda, são calculados os parâmetros da distribuição de probabilidade normal das medições. A série observada é truncada e reavaliada sua média e desvio padrão, processo que se repete até atingir uma quantidade mínima de leituras na amostra suficiente para a determinação de tais parâmetros. O comportamento esperado dos parâmetros em relação ao nível de truncamento é definido teoricamente, e uma divergência abrupta nele é interpretada como um vazamento. Também analisando essa divergência é estimada a ordem de grandeza das perdas. Segundo os autores, para aplicação da técnica é preciso contar com medições de alta frequência, da ordem de um valor cada 5s a 1s, e uma área monitorada com menos de mil residências.

Usando ferramentas mais complexas para análise dos sinais de vazão num setor de rede, Mounce, Boxal e Machell (2010) propõem um método para detecção de vazamentos e outras anormalidades em tempo real, usando inteligência artificial, cujas bases foram apresentadas em publicações anteriores (MOUNCE et al., 2006, 2007, 2008 apud (ROMANO; KAPELAN; SAVIC, 2009). Dados de vazão coletados na rede com intervalos de 15 minutos, durante os três meses imediatamente anteriores à análise são submetidos a um pré-processamento, para serem utilizados no treinamento de uma rede neural artificial (RNA). Posteriormente é construído o modelo de densidade de probabilidade para as vazões futuras produzidas pela RNA. Séries atualizadas de valores observados são comparadas em tempo real com as previsões para períodos de durações variáveis; com um sistema de inferência difusa é gerado um alerta quando detectadas observações anormais. Baseado na densidade de probabilidades, para cada evento é associado um intervalo de confiança que permite classificar e priorizar os alarmes. O método foi testado para uma rede de distribuição de água potável do Reino Unido com sucesso.

Por outro lado, o método apresentado por K. Aksela, M. Askela e Vahala (2009) está baseado nos mapas auto-organizados (SOM, do inglês *Self Organizing Maps*), que correspondem a um tipo de RNA treinada usando aprendizado não supervisionado para o agrupamento de um conjunto de entradas e que produz sua representação com menor dimensionalidade, que são os chamados mapas. As informações utilizadas para a construção do modelo de vazamentos são sinais de vazão e registros de vazamentos conhecidos, num setor de rede real. Uma fórmula empírica de vazamento em função da distância ao sensor e do tempo é definida para o caso estudado. As entradas do SOM são os sinais de vazão, depois do pré-processamento, com base semanal, correspondentes a todo o período de dados. A função de vazamento empírica é incorporada, resultando na construção de mapas dos padrões de vazão usados para indicar o valor da função de vazamento. Depois do processo de treinamento antes descrito, os “mapas” criados permitem identificar um vazamento.

Ainda dentro das técnicas baseadas nos registros de vazão em setores monitorados da rede, também está a análise de componentes principais (PCA, do inglês *Principal Component Analysis*) para detecção de vazamentos (PALAU; ARREGUI; CARLOS, 2012). A técnica de PCA permite, com uma transformação ortogonal, reduzir uma matriz de alta dimensionalidade como a de vazões num setor durante um determinado período, para uma matriz de dimensionalidade menor na qual toda a variabilidade dos dados é expressa num conjunto de poucas variáveis linearmente descorrelacionadas. Assim, é construído um modelo estatístico que representa o comportamento da rede baseado exclusivamente nos sinais adquiridos. Quando um vazamento se incorpora à vazão, os valores das componentes mudam, e fazendo uso de técnicas estatísticas de controle é possível a identificação de anomalias, incluindo início de vazamentos.

Incorporando a análise de sinais de pressão juntamente aos de vazão para a detecção de vazamentos, foi proposto o uso de filtros Kalman (YE; PH; FENNER, 2011), um método estatístico que utiliza trechos conhecidos de séries temporais para produzir estimativas do estado do sistema, no caso as pressões e vazões. Os filtros Kalman fornecem uma caracterização estatística de sistemas dinâmicos, como o sinal temporal de vazão ou pressão na rede, considerando a influência dos valores passados na distribuição de probabilidades em um processo de aprendizado incremental. Assim, os filtros Kalman são aplicados para estimar os parâmetros hidráulicos (pressão e vazão) na rede em condições normais, e um vazamento pode ser detectado analisando as parcelas residuais do filtro, que correspondem à diferença entre valores estimados e medidos.

Usando também o monitoramento de pressões e vazões na rede durante a operação, foi desenvolvido um método baseado em inferência bayesiana (ROMANO; KAPELAN; SAVIC, 2009, 2010). O método utiliza dados de vazão na entrada do setor e de pressões em um ou mais pontos internos a ele, adquiridos a uma taxa de 1 amostra cada 15 minutos. A esses sinais aplica-se a remoção do ruído usando a transformada *wavelet*, que se apresentam como uma boa ferramenta para o propósito de separar o ruído de outras características não permanentes do sinal. Os sinais sem ruído do último período de 24 horas constituem a informação de treinamento para um modelo de previsão de curto prazo, usando o *Group Method of Data Handling*, técnica estatística que utiliza todas as informações disponíveis sem a necessidade de especificar a forma do modelo para a previsão. Valores atualizados adquiridos na rede são comparados com os preditos pelo modelo, identificando as discrepâncias, e depois metodologias de controle estatístico de processos são usadas para incrementar a confiabilidade da hipótese de vazamento. As informações assim obtidas são entregues a um sistema de inferência Bayesiana para a avaliação da probabilidade de ocorrência de um vazamento e a geração de alarmes.

A mesma equipe de pesquisa fez melhorias na proposta (ROMANO; KAPELAN; SAVIC, 2012), desenvolvendo um sistema modular que incorpora ferramentas adicionais em cada um dos passos do processo de detecção. Depois do pré-processamento e filtragem do ruído, a análise de discrepâncias é feita com uma Rede Neural Artificial (RNA). A RNA é treinada com os sinais coletados em períodos anteriores, e recebe como entradas dados atualizados para produzir como saída os valores futuros preditos, que são comparados com os dados adquiridos da rede e calculadas as discrepâncias. Adicionalmente avaliações estatísticas dos desvios no sinal e da sua tendência são realizadas para gerar evidência complementar sobre a ocorrência de vazamento. As evidências geradas nas análises de discrepâncias, dos desvios e das tendências são incorporadas a um sistema de inferência Bayesiana que estima a probabilidade da ocorrência de um vazamento, combina as informações de diferentes sensores e finalmente gera os alarmes de vazamento quando um limiar é atingido. A técnica tem sido testada em um setor de rede real (ROMANO; KAPELAN; SAVIC, 2012).

As técnicas baseadas na análise de sinais durante a operação da rede (AS_O), apesar do curto tempo desde seu surgimento, têm se mostrado promissoras. Porém, não têm sido publicadas suas aplicações efetivas a sistemas reais, possivelmente pelo tempo que naturalmente leva a transferência de tecnologia para a indústria, mas também por características das próprias técnicas que precisariam ser melhoradas para sua implementação generalizada. A aplicação de

técnicas de inteligência artificial na análise dos sinais nas técnicas AS_O é ainda inovadora, sendo que, a cada nova pesquisa com resultados positivos, novas possibilidades de melhoria são abertas. Reconhecendo que os objetivos práticos do seu desenvolvimento não têm sido satisfatoriamente atingidos, há um grande interesse por continuar explorando esta abordagem.

4. CONCEITOS DE ANÁLISE DE SINAIS E APRENDIZADO DE MÁQUINA

A metodologia proposta para detecção de vazamentos, com o suporte de um sistema de reconhecimento de padrões, têm suas bases em disciplinas diferentes da engenharia hidráulica, como o processamento de sinais e o aprendizado de máquina. Com o objetivo de apresentar os conceitos básicos aplicados no trabalho para leitores não familiarizados com esses temas, neste capítulo são resumidas definições e explicações simples, que poderão ser aprofundadas na bibliografia citada.

4.1.VARIÁVEIS DO SISTEMA COMO SINAIS

Um sinal pode ser definido como qualquer ente físico que varia ao longo do tempo, espaço ou outra variável independente (ou várias delas) (PROAKIS; MANOLAKIS, 2007). Todo sinal pode ser representado matematicamente como função de uma ou diversas variáveis independentes. Segundo a definição anterior tanto a pressão quanto a vazão em pontos de uma rede podem ser estudados como sinais que, para as análises de interesse, são considerados independentes entre si, de forma a considerá-los como sinais unidimensionais, variáveis no tempo somente.

É entendendo as variáveis do sistema, pressão e vazão, como sinais, que foram neste trabalho exploradas metodologias de processamento de sinais. O termo “processamento de sinais” é muito abrangente, pois considera todos os sistemas (físicos ou lógicos) que realizam operações sobre o sinal, seja com o fim de produzir um sinal alterado ou de extrair informações de interesse, neste último caso com o nome mais apropriado de “análise de sinais”.

Os sinais podem ser classificados em dois tipos básicos: analógicos e digitais. Se a variável independente é contínua e os valores da variável dependente estão também numa faixa contínua, o sinal é analógico. Se tanto a variável independente quanto a variável dependente são discretas, então o sinal é dito digital. A grande maioria de sinais no mundo físico, incluindo a pressão e a vazão, é analógica: por exemplo, em todo instante de tempo sem interrupção existe uma pressão, e sua magnitude instantânea pode assumir infinitos valores. É possível o processamento de sinais analógicos, mas os métodos mais avançados de análise e

processamento que envolvem o uso de processadores são aplicáveis somente a sinais digitais, razão pela qual a conversão dos sinais se faz necessária.

A conversão de sinais analógicos em digitais é definida por três processos: amostragem, quantização e codificação. A amostragem refere-se à extração de valores do sinal em intervalos discretos de tempo, a quantização à aproximação dos valores lidos a valores discretos, e a codificação à expressão dos valores no formato adequado para transmissão das informações, isto é, uma sequência de bits (PROAKIS; MANOLAKIS, 2007). Depois da conversão para sinais digitais, as variáveis de interesse estarão representadas por uma série de valores, um para cada instante de tempo em intervalos fixos, sendo que cada valor será arredondado para o respectivo valor discreto mais próximo, e essas informações serão transmitidas para seu processamento.

Para o presente trabalho de pesquisa, e como apresentado na seção 5.2, as pressões e vazões na rede foram medidas pelos transdutores e hidrômetros, e transmitidas para uma placa de aquisição de dados que controla o processo de conversão e transmissão do sinal digital para o microcomputador.

4.2. TRANSFORMAÇÕES DE DOMÍNIO

A pressão e vazão na rede são neste trabalho, segundo explicado na seção anterior, sinais digitais variáveis no tempo, isto é, sinais temporais discretos. O sinal é composto por séries de valores instantâneos, e pelo processo de digitalização usado, os intervalos de tempo entre valores sucessivos são constantes. De acordo com essa representação temporal, as mudanças nos valores correspondem a mudanças no estado do sistema observado, em função do tempo.

Quando um sinal é analisado nessa representação, como uma sequência de valores no tempo, fala-se de análise no domínio do tempo. Com o sinal no domínio do tempo podem ser feitas transformações no sinal tais como a padronização (seção 6.1), podem ser calculadas características como energia, potência, entropia e outras (seção 6.2), podem realizar-se operações como adição, multiplicação ou mudança de escala, dentre muitas outras opções de análise e processamento de sinais.

Sem desconhecer a variedade e importância das metodologias de processamento de sinais no domínio do tempo, em muitos casos a representação temporal é insuficiente e outras formas

de indexação podem mostrar-se superiores, como a transformação para o domínio da frequência (MEYER, 1999).

4.2.1. Domínio da frequência e Transformada de Fourier

A definição original da frequência, vinda da física, relaciona-se com o movimento periódico, em particular com a oscilação harmônica, comumente descrita por funções senoidais. Nesse caso, a frequência refere-se à mudança angular por unidade do tempo na função senoidal, ou também à quantidade de ciclos por segundo (PROAKIS; MANOLAKIS, 2007).

Seja considerado um sinal temporal discreto periódico (no qual as mesmas sequências de valores repetem-se a cada certo número de intervalos no tempo), por exemplo, um sinal senoidal. Nesse sinal simples, o período é definido como o tempo transcorrido para a repetição da sequência de valores (ciclos), e a frequência é o inverso do período, ou o número de ciclos numa unidade de tempo. Portanto, esse sinal senoidal simples terá um único valor de frequência representativo, comumente medido em unidade de $s^{-1} = \text{Hz}$. Se o sinal não fosse simplesmente senoidal, mas composto pelo somatório de várias funções senoidais de diferentes períodos, ao invés de um único valor o sinal teria múltiplos valores de frequência presentes. Um sinal mais complexo é composto por uma faixa muito ampla de frequências diferentes. A análise dessas frequências presentes no sinal, ao invés dos valores em função do tempo, corresponde à análise no domínio da frequência.

A análise no domínio da frequência é indispensável na área de processamento de sinais. A descrição em termos da frequência pode ser a base de uma melhor compreensão do fenômeno que dá origem aos sinais, sendo que oferece um complemento essencial à descrição exclusivamente temporal (MEYER, 1999). A ferramenta padrão para análise da frequência é a transformada de Fourier, cuja ideia básica é que qualquer função pode ser representada pelo somatório de funções senoidais de diferentes frequências e fases. Assim, usando a transformada de Fourier obtêm-se os componentes relativos às diferentes frequências no sinal, construindo uma representação diferente e independente do eixo temporal. O processo é reversível, pois utilizando a transformada inversa de Fourier, pode-se passar do domínio da frequência para o do tempo.

A transformada de Fourier funciona muito bem no processamento de sinais estacionários (invariantes no tempo), mas tem limitações quando são analisados sinais não estacionários ou problemas envolvendo singularidades no sinal como mudanças localizadas no tempo

(MALLAT, 1999). A principal limitação deve-se a que com esse tipo de transformação é possível saber quais frequências compõem um sinal, mas sem nenhuma indicação do momento (ou valor no eixo temporal original) no qual essas frequências se apresentam.

4.2.2. Transformada *Wavelet* e análise tempo-frequência

Reconhecendo que as representações em tempo e em frequência são necessárias, mas insuficientes, foram desenvolvidos métodos de análise que efetivamente fornecessem uma descrição conjunta do sinal, nos dois domínios simultaneamente, e são as chamadas análises tempo-frequência. Uma das principais aproximações para o problema é a decomposição dos sinais a respeito de uma família de sinais ou funções elementares, as quais são bem localizadas tanto em tempo quanto em frequência (diferentemente da transformada de Fourier tradicional). Os métodos mais importantes que seguem essa aproximação são a decomposição Gabor (relacionada com a transformada de Fourier de curta duração) e a decomposição *wavelet* (MEYER, 1999).

As *wavelets* são simplesmente funções que satisfazem certas condições matemáticas e que são usadas para a representação de dados ou de outras funções (GRAPS, 1995). As funções *wavelet* devem ter media zero, localização no tempo e energia finita. Essas características são comuns a inúmeras funções matemáticas, mas na prática são aplicadas algumas funções amplamente estudadas e validadas. Neste trabalho será utilizada uma das funções mais comuns, a função Daubechies com dois “momentos zero”, ou db2, apresentada na Figura 3. A escolha desta função com dois momentos, que corresponde com filtros de suporte quatro, esta relacionada à sua capacidade para análises onde a resolução temporal é importante (ADDISON, 2002). Considerando também que o fenômeno analisado (o vazamento) esta relacionado com uma descontinuidade na primeira derivada do sinal não se considera conveniente o uso de funções com maior suporte.

Figura 3. Função *wavelet* Daubechies, db2

As funções *wavelet* são combinadas, mudando sua escala e posição, para transformar o sinal em análise para uma representação mais útil. A mudança de escala, que equivale à expansão ou contração da função, está também indiretamente relacionada com as frequências presentes no sinal, enquanto a posição refere-se ao eixo temporal original do sinal. Assim, a transformada *wavelet* permite analisar as frequências presentes no sinal sem perder as informações referentes ao sinal no tempo. A transformada *wavelet* discreta de um sinal é definida como segue:

$$(Wx)_{(m,n)} = \int_{-\infty}^{\infty} x(t)\psi_{m,n}(t)dt = \int_{-\infty}^{\infty} x(t) \frac{1}{\sqrt{2^m}} \psi\left(\frac{t - n2^m}{2^m}\right) dt$$

Sendo x = sinal original, em função do tempo

$W_{x_{m,n}}$ = Transformada *wavelet* do sinal x , na escala m e posição n

$\psi_{m,n}(t)$ = Função *wavelet*, em função do tempo, na escala m e posição n

A característica de localização das funções elementares e a possibilidade de análise em múltiplas escalas ou resoluções são vantagens importantes da transformada *wavelet*, que permite que as informações sejam adequadamente separadas, o que é importante para aplicações como a identificação de singularidades ou a eliminação de ruído em sinais.

Matematicamente, a transformada *wavelet* é também a convolução da função *wavelet* com o sinal sob análise¹, e é possível entender essa convolução como um processo de filtragem do sinal. O termo “filtragem”, que em geral refere-se ao processo de discriminação de objetos segundo seus atributos, na análise de sinais é usado comumente em referência à discriminação das componentes de frequência diferentes presentes no sinal de entrada. Assim, um sinal filtrado contém somente algumas das frequências presentes originalmente no sinal, enquanto outras são descartadas.

Assim, relacionando a transformada *wavelet* com um processo de filtragem, é possível entender a transformação em cada nível (escala) como uma filtragem dupla, onde do sinal original são extraídas uma versão contendo só as frequências baixas nele presentes e outra versão contendo só as frequências altas.

¹ A convolução é um operador linear que transforma duas funções em uma terceira função, no caso a convolução da função *wavelet* e o sinal dá origem ao sinal transformado.

4.3.DECOMPOSIÇÃO WAVELET DISCRETA E PACKET

Na mesma linha de raciocínio da interpretação da transformação *wavelet* como processo de filtragem, a decomposição *wavelet* discreta de um sinal pode ser entendida como a representação dele em diferentes escalas, relacionadas com as frequências. Essa nova representação é obtida decompondo o sinal em duas partes, uma correspondente às baixas frequências e representada por um vetor de coeficientes de aproximação, enquanto a outra parte correspondente às altas frequências e representada por um vetor de coeficientes de detalhe. Assim, o sinal original fica representado por dois componentes depois de decomposto no nível $m = 1$. O processo de decomposição em aproximação e detalhe é executado sucessivamente, com base na aproximação de nível imediatamente anterior, como mostrado na Figura 4. O número máximo de níveis de decomposição, M , depende do tamanho original do sinal, $N = 2^M$. Decompondo até um nível m qualquer ($1 \leq m \leq M$), o sinal fica representado por $m+1$ componentes, sendo um vetor de coeficientes da aproximação, e m vetores de coeficientes dos detalhes.

Figura 4. Esquema da decomposição *wavelet* discreta

A quantidade de elementos nos vetores de aproximação e detalhes muda de forma tal que em todo nível m o total de elementos de todas as componentes é igual a N . Adicionalmente, cada uma das componentes corresponde a uma representação da duração completa do sinal. Lembrando também que cada decomposição produz uma divisão mais fina da faixa de frequências contidas em cada componente, entende-se que níveis de decomposição (m) maiores implicam uma maior resolução na frequência, mas menor resolução temporal,

enquanto níveis m menores fornecem melhor resolução temporal, mas pobre resolução de frequências. A seleção do melhor nível de decomposição varia de um problema para outro, sendo necessário avaliar conjuntamente o ganho e o custo do cálculo. Essa seleção para o presente estudo é apresentada na seção 7.1.

A decomposição *wavelet packet* é uma generalização da decomposição *wavelet* discreta. Enquanto na decomposição *wavelet* discreta os componentes de detalhe passam inalterados para os níveis de decomposição seguintes, na decomposição *wavelet packet* em cada nível tanto os componentes de aproximação quanto os de detalhe são decompostos sucessivamente. Assim, para um nível de decomposição m , o sinal fica representado por um vetor de aproximação e $2^m - 1$ vetores de detalhes, como mostrado na Figura 5.

Figura 5. Esquema da decomposição *wavelet packet*

Lembrando que os vetores de aproximação correspondem às frequências mais baixas enquanto os de detalhes às frequências mais altas, ao aplicar a decomposição *wavelet packet* ao invés da *wavelet* discreta, obtém-se informações mais detalhadas de sub-faixas de frequência mais altas.

A geração e o processamento dessa maior quantidade de informações implicam num alto custo computacional, e é necessário avaliar o benefício de sua aplicação a cada problema, como é feito na seção 7.1.

4.4. APRENDIZADO DE MÁQUINA

Comumente a resolução de problemas complexos, envolvendo grande volume de dados e para os quais não é conhecida uma solução algorítmica é definida, deve ser abordada computacionalmente, requerendo sistemas que além de executar operações conhecidas sejam capazes de criar por si próprios, a partir de experiências passadas, uma hipótese ou função capaz de resolver o problema. Entra-se assim no campo do aprendizado de máquina, pela qual os computadores são programados para melhorar o desempenho por meio da experiência, obtendo conclusões genéricas a partir de conjuntos de exemplos, i.e. com um processo de indução (FACELI et al., 2011).

Assim, o aprendizado indutivo é realizado pela análise de exemplos obtidos com anterioridade ao processo de aprendizado. Segundo as informações anteriores fornecidas ao processo de aprendizado, este pode ser supervisionado ou não supervisionado. No aprendizado supervisionado para cada um dos exemplos é fornecida também a resposta ou saída correspondente, enquanto no aprendizado não supervisionado o próprio algoritmo deverá tentar agrupar e a interpretação desse agrupamento usualmente requer uma análise posterior.

O problema tratado neste trabalho usando aprendizado de máquina é a distinção de sinais de pressão ou vazão segundo a condição na qual foram adquiridos: durante a ocorrência de um vazamento ou não. Como a resposta esperada é se o sinal sob análise pertence a uma ou à outra classe, trata-se de um problema de classificação. A classificação é um caso particular de aprendizado supervisionado onde aos exemplos estão associados rótulos de classe pré-determinado. Neste trabalho, existem somente duas classes (corresponde com a situação de vazamento ou não), portanto dois valores para o rótulo de classe, formando um problema de classificação binária. A metodologia de abordagem do problema é apresentada no capítulo 6.

O processo de aprendizado indutivo, durante o qual são obtidos os elementos necessários para extrair generalizações a partir dos exemplos rotulados, também costuma ser chamado de treinamento, termo que será usado neste trabalho.

5. AQUISIÇÃO DE SINAIS EXPERIMENTAIS EM LABORATÓRIO

O horizonte da detecção de vazamentos abordada no presente trabalho refere-se às redes de distribuição de água para abastecimento, sistemas complexos e muito pouco analisados sob a perspectiva da análise de sinais aqui proposta, como mostrado na seção 3.2. Levando em conta a novidade das metodologias propostas e a necessidade de encarar o problema primeiramente sob condições conhecidas e controladas, decidiu-se trabalhar com um sistema de laboratório que será descrito a seguir, juntamente com as condições de experimentação adotadas para a geração das bases de dados utilizadas em todas as análises posteriores.

5.1.MONTAGEM EXPERIMENTAL

O circuito hidráulico experimental é um sistema para estudo sobre uma rede de condutos forçados, com monitoramento automático de pressões e vazões, localizado no laboratório de Hidráulica Ambiental do Centro de Recursos Hídricos e Ecologia Aplicada (CRHEA). O sistema completo é composto pela rede de tubulações em policloreto de vinila (PVC) desenvolvida verticalmente, o sistema de alimentação e o de aquisição de dados, cujo esquema é apresentado na Figura 6.

O circuito hidráulico é formado por um conjunto de tubulações de PVC de 53, 75 e 101 mm de diâmetro interno com comprimento total de 203 m, além de todos os acessórios hidráulicos necessários (tês, cotovelos, reduções e ampliações, entre outros). O sistema possui também 37 válvulas (18 válvulas esfera e 19 válvulas de gaveta) para a modificação da topologia do sistema e o controle de vazões, além de uma válvula automática para geração de fenômeno transitório no sistema, e nove válvulas para saída de vazão que permitem a modelação de vazamentos nos nós.

Para a alimentação do sistema hidráulico é utilizada uma estação de bombeamento, com duas bombas centrifugas em paralelo, de potência 1 cv e 5 cv respectivamente, e com uma válvula de retenção a jusante delas. O bombeamento se dá a partir de um reservatório com nível de 5 m, ao qual também retorna a vazão depois de circular pela rede, constituindo um circuito fechado.

Figura 6. Esquema do circuito hidráulico experimental

5.2.AQUISIÇÃO DE DADOS

A aquisição de dados é feita por 15 transdutores de pressão com escalas de até 100, 200 e 300 psi, precisão de 0,25% do fundo de escala, e saída em corrente na faixa de 4 a 20 mA. Para a medição de vazões de saída do circuito pelas válvulas de simulação de vazamentos são usados 9 medidores tipo hidrômetro com sensor eletrônico incorporado. Três medidores eletromagnéticos de vazão de 100 mm de diâmetro, com escala de 0 a 50 m³/h, estão localizados nas tubulações de entrada, de saída e uma intermediária. A localização e nomenclatura dos medidores são apresentadas na Figura 7.

Todos os sensores estão ligados a conversores analógicos para ajustar os diferentes tipos de sinal que alimentam uma placa de aquisição de dados. A placa está comunicada com um microcomputador no qual um *software* especificamente desenvolvido para o projeto permite controlar a aquisição. No *software* também são incorporadas as equações de calibração de cada um dos sensores e é possível fazer acompanhamento. É possível acompanhar os valores das variáveis adquiridos a partir de cada sensor, com atualização a cada segundo, graficamente. Para a gravação dos dados durante experimentos é possível escolher os sensores a serem considerados, a taxa de aquisição (entre 1 amostra/minuto e 12000 amostras/minuto), e o tempo total de aquisição (de 1 minuto até 48 horas). Também através do sistema as bombas e a válvula automática podem ser acionadas usando o *software*. Na Figura 8 é apresentada uma vista da interface visual do *software*.

Figura 7. Localização dos sensores no circuito hidráulico experimental

Figura 8. Interface do *software* de aquisição de dados

5.3.RESUMO DOS EXPERIMENTOS

Para a construção e teste do sistema de reconhecimento de padrões foi usada uma base de dados de exemplos adquiridos junto ao circuito hidráulico experimental. Com os 15 sensores de pressão e 3 de vazão simultaneamente, foram adquiridos exemplos de duas classes: sem vazamento (SEM) e durante o início de um vazamento (COM). No total, para cada um dos 18 sensores conta-se com 620 exemplos, sendo 310 para cada uma das classes.

Os exemplos da classe SEM correspondem a medições durante operação do circuito com vazão (Q) permanente, mas não igual para todos os exemplos. Foram feitas medições contínuas com taxa de 4 amostras/segundo, e posteriormente foram segmentados sinais de exemplo (ver seção 6.1), com duração de um minuto ou 240 valores cada. O ensaio foi repetido usando diferentes valores de vazão no circuito (Q) até conseguir completar 310 exemplos da classe. Na Figura 9 são apresentados três exemplos da classe SEM, correspondentes a três diferentes valores de Q.

Figura 9. Três exemplos de sinais da classe SEM adquiridos com o sensor P02 para três diferentes vazões no circuito (Q).

Por sua vez, os exemplos da classe COM foram adquiridos durante experimentos nos quais um registro de simulação de vazamento (V1 na Figura 7) era aberto rapidamente. Na Figura 10 são apresentados três exemplos da classe COM nos quais é possível identificar visualmente a queda de pressão devida ao vazamento. Salienta-se que essa identificação é difícil para a maior parte dos exemplos.

Figura 10. Três exemplos de sinais da classe COM adquiridos com o sensor P02 para três diferentes vazões no circuito (Q) e diferentes magnitudes de vazamento (V).

Para os exemplos COM, mantendo a taxa de 4 amostras/segundo e mediante processo de segmentação foram obtidos 310 sinais de exemplo, com duração de 1 minuto ou 240 valores, cada um deles relativo aos momentos antes, durante e depois da abertura do registro de simulação de vazamento (realizando, para tanto, 310 ensaios). Nos experimentos a vazão do circuito (Q) foi controlada pela abertura manual da válvula de saída enquanto a magnitude de vazamento (V) dependeu da abertura do registro para sua simulação. Assim, buscou-se que as vazões tanto Q quanto V apresentassem variações de um exemplo para outro, para representar as diferentes condições de operação do circuito. Com base nos valores medidos pelos sensores foram calculadas as relações V/Q resultantes para cada exemplo, cujo resumo para o total dos 310 exemplos apresenta-se na Tabela 2.

Tabela 2. Relações V/Q dos exemplos COM

V/Q	< 3%	3% - 4%	4% - 5%	5% - 6%	6% - 7%	7% - 8%	8% - 9%	9% - 10%	10% - 11%	> 11%
Quant. de exemplos	10	33	51	39	30	36	29	35	32	15

6. SISTEMA DE RECONHECIMENTO DE PADRÕES

Segundo foi explicado anteriormente, no presente trabalho o objetivo do processamento dos sinais de pressão e de vazão em uma rede sob pressão é a identificação de singularidades no sinal que possam indicar a ocorrência de um vazamento na rede.

Provavelmente o jeito mais rudimentar de abordar a tarefa seria examinar uma representação gráfica do sinal e determinar se há ou não uma anormalidade, com base no conhecimento prévio, capacidade e até intuição. Claramente, essa metodologia tem grandes limitações, é praticamente impossível usá-la para processar uma quantidade importante de informações. Ainda seus resultados serão diferentes, por exemplo, segundo a pessoa que faz a avaliação, a escala e a forma de representação do sinal, dentre muitos outros fatores, fazendo com que grande quantidade de singularidades seja despercebida e a taxa de falsos alarmes seja grande demais.

Enfrentando a necessidade de automatizar o processo de reconhecimento das singularidades no sinal, são propostas as metodologias conhecidas como reconhecimento de padrões, uma grande área dentro do processamento digital de sinais. O reconhecimento de padrões é basicamente o processo de análise automática de uma entrada (na forma de uma representação digital) para relacioná-la com uma categoria de saída. Como é deduzido do próprio nome, um sistema de reconhecimento de padrões visa definir se um sinal de entrada corresponde ou não (e com qual chance de acerto) a um padrão dado. Diversos campos das ciências básicas e da engenharia fazem uso do reconhecimento de padrões, aplicando-o para o processamento de imagens, músicas, sinais de voz, imagens diagnósticas e todo tipo de sinais.

Um sistema de reconhecimento de padrões tem uma estrutura básica que compreende várias etapas; um novo sinal de entrada pode passar (ou não) por processo de segmentação, depois passa pela extração de características e finalmente é classificado. Segundo indicado na seção 4.4, a classificação é um problema de aprendizado de máquina supervisionado no qual é utilizada a informação de exemplos acertadamente classificados para possibilitar que um novo sinal seja atribuído a uma das classes. Portanto a resposta que entrega o sistema é o nome da classe à qual pertence o sinal de entrada dado. Cada uma das etapas, mostradas na Figura 11, será brevemente descrita em este documento.

Figura 11. Esquema básico das etapas de um sistema de reconhecimento de padrões

A construção de um sistema de reconhecimento de padrões é fortemente dependente do tipo de sinal e do objetivo para o qual foi construído, existindo grande diversidade de metodologias aplicáveis a cada uma das etapas. Deve ser lembrado que estes sistemas não existem isoladamente, mas como parte de procedimentos mais completos que integram a aquisição dos sinais de entrada, e posteriormente o uso das saídas como parte de um sistema de suporte à decisão.

Focando o problema em estudo, os sinais de entrada correspondem a pressões e vazões adquiridas no sistema, o processo de segmentação visa isolar trechos de curta duração, a extração de características é aplicada para obter uma representação padronizada de cada um, e um classificador binário define se o sinal pertence à mesma classe dos sinais adquiridos durante a ocorrência de um vazamento (COM) ou à dos adquiridos na ausência deles (SEM).

6.1.SEGMENTAÇÃO E PADRONIZAÇÃO

Pela natureza do sinal coletado, é necessário o processamento de grande quantidade de dados, e não é interessante realizar a análise somente para um sinal de grande comprimento. Por exemplo, se o sinal for adquirido durante várias horas, não é interessante definir somente um conjunto de valores característicos (extração de características) que represente todo o comprimento do sinal porque poderia mascarar as mudanças que aconteceram em algum momento durante o intervalo de aquisição e que se deseja sejam reconhecidas. Para solucionar esse problema são propostas as técnicas de segmentação (DUDA; HART; STORK, 2001).

De forma geral, as técnicas de segmentação permitem modificar um sinal de entrada de tamanho grande ou que envolve varias repetições do fenômeno a estudar, para obter sinais

divididos em unidades de análise mais adequadas, ou segmentos. No caso de sinais temporais de grande tamanho como os de pressão e vazão adquiridos, essa segmentação pode ser feita definindo como unidades de análise intervalos de duração curta, o que pode conseguir-se com a metodologia de janelas deslizantes.

A técnica de janelas deslizantes consiste em definir sucessivamente subconjuntos de dados consecutivos do sinal (ou equivalentemente intervalos de tempo), sendo que cada novo subconjunto (ou intervalo) desconsidera os primeiros dados do subconjunto anterior e incorpora os próximos elementos do sinal; desta forma, a segmentação é feita “percorrendo” o comprimento total do sinal, extraindo a quantidade necessária de segmentos ou subconjuntos de tamanho fixo, ficando assim explicado o termo de janela deslizante. As etapas do sistema de reconhecimento de padrões podem então ser aplicadas a cada subconjunto ou segmento.

O uso da janela deslizante como técnica para segmentação tem diferentes variações, entre elas a “forma da janela”, ou a função pela qual são lidos os dados do sinal para construir cada segmento. O tipo de janela mais simples é a janela retangular, que pode ser representado como:

$$X_i = S \cdot j$$

$$j_m = \begin{cases} 1 & \text{se } m \in [\text{posicao}_i, \text{posicao}_i + N] \\ 0 & \text{em outro caso} \end{cases}$$

Sendo S = Sinal original, tamanho M

X_i = i -ésimo segmento extraído do sinal original, tamanho N ($N < M$)

posicao_i = posição do início do i -ésimo segmento no sinal S

$m = 1, 2 \dots M$

$i = 1, 2 \dots T$; T = número de segmentos de tamanho N que podem ser extraídos de X

Segundo os objetivos da análise, as propriedades do sinal e das outras etapas do sistema de reconhecimento, os efeitos de borda produzidos ao restringir as análises a um conjunto finito e curto de dados podem ser importantes. Para diminuir esse efeito indesejável cada subconjunto pode ser constituído por dados extraídos com funções j mais complexas do que a função retangular antes descrita, por exemplo, curvas Gaussianas ou outras com fatores tendendo a zero nos extremos. Para a análise dos sinais de pressão e vazão na rede, usando o conjunto de características apresentado na seção seguinte, é estimado desnecessário o uso de janelas com

formas complexas. Portanto, é definido o uso da janela retangular como a metodologia de segmentação para todos os casos analisados.

O tamanho de segmentos mais conveniente pode variar muito de um para outro problema e não existem regras definidas para a sua escolha. Atendendo às necessidades específicas dos dados adquiridos experimentalmente, foi definido usar segmentos com 240 pontos, o que significa duração de 1 minuto com taxa de amostragem de 4 amostras/segundo. A técnica foi adaptada às condições de execução dos ensaios, de forma a garantir a adequada distinção entre segmentos relativos às classes COM e SEM. Cada um dos segmentos será considerado nas análises seguintes como um exemplo isolado, passando a usar o termo sinal para cada um dos segmentos extraídos do sinal adquirido.

Os valores de pressão e vazão medidos são dependentes das condições do circuito: nível no reservatório, operação das bombas, vazão no circuito e, no caso dos exemplos COM, da vazão de saída pelo vazamento. É necessário então considerar um dos problemas básicos para o correto funcionamento de um sistema de reconhecimento de padrões: a generalização. Sendo que o objetivo é identificar se um novo sinal adquirido inclui ou não o início de um vazamento, o projeto do sistema não pode ser dependente de algumas condições no circuito, e sim ser capaz de fornecer resultados satisfatórios com sinais adquiridos sob uma nova condição, isto é, ter capacidade de generalização.

As diferenças nas grandezas medidas não estão relacionadas com o fenômeno que queremos reconhecer, o vazamento. Assim, visando à capacidade de generalização do sistema de reconhecimento, essas diferenças podem ser desconsideradas. Para tanto, todos os sinais utilizados (depois do processo de segmentação) são padronizados, segundo a expressão seguinte:

$$x_i = \frac{y_i - \bar{y}}{|y|_{\max} - \bar{y}}$$

Sendo $X = \{x_1, x_2, \dots, x_N\}$ = Sinal padronizado de tamanho N

$Y = \{y_1, y_2, \dots, y_N\}$ = Sinal sem padronizar de tamanho N

$\bar{y}$ = média do sinal sem padronizar

$|y|_{\max}$ = máximo valor absoluto dos elementos do sinal sem padronizar

6.2.EXTRAÇÃO DE CARACTERÍSTICAS

Cada um dos sinais usados na pesquisa é representado por um vetor de 240 valores sequenciais de pressão ou vazão, que podem estar afetados por ruído, variações originadas na medição, transmissão, ou ainda por fenômenos físicos como micro turbulências. Pela conformação dos exemplos, o tamanho de todos eles é igual, mas em condições menos controladas são comuns sinais com diferentes durações ou taxa de amostragem, dados faltantes ou valores errados que precisam ser desconsiderados, de forma que o sinal ou vetor de entrada comumente tem tamanho variável e grande.

Assim, a análise dos valores brutos do sinal adquirido pode ser prejudicada pela presença de perturbações, a variabilidade no tamanho dos sinais, e o alto custo computacional da análise com séries longas. Em resposta a esses problemas, diversas metodologias de processamento de sinais fazem uso da extração de características. A extração de características permite analisar, ao invés do grande vetor de valores do sinal, características que dele derivam e que podem ser representadas em um vetor muito menor, que ainda mantenha a capacidade de distinguir a classe a que um exemplo pertence.

Para maior clareza no conceito de extração de características, e a maneira de exemplo, na Figura 12 apresentam-se dois sinais de pressão de 1 minuto ou 240 pontos adquiridos, cada um pertencendo a uma classe diferente (SEM e COM). É importante salientar que a diferença na ordem de grandeza desses sinais é devida à diferença na vazão no sistema durante a sua aquisição. Já uma diferença no formato (ou comportamento) que permita uma identificação visual do início de um vazamento não é clara a partir desses sinais diretamente. Depois de normalizar os valores das séries, as quatro características consideradas foram extraídas desses dois sinais de exemplo, formando para cada um deles um vetor de sete elementos², cujos valores estão nos eixos verticais da Figura 13.

Ao comparar os valores das características resultantes é possível identificar uma diferença entre o exemplo COM e SEM de forma mais convincente que comparando seus comportamentos na Figura 12. Claramente, no presente trabalho de pesquisa, a comparação não é feita só com dois sinais, como no exemplo anterior, e sim com todos os dados adquiridos, obtendo os resultados apresentados na seção correspondente.

² Os valores das características extraídas são vetores de um ou mais elementos, é por isso que ao extrair as quatro características (explicadas na seção 6.2) o resultado é um vetor de sete elementos.

Figura 12 Exemplo de dois sinais de pressão adquiridos.

Figura 13 Características extraídas dos dois sinais de pressão.

Existe uma grande diversidade de características que podem representar o sinal, que podem ser julgadas mais ou menos importantes, considerando o tipo de sinal e os objetivos da análise. Para identificar quais as características que melhor poderiam ser adaptadas ao sistema de reconhecimento de padrões para detecção de vazamentos, foi explorada a extração de quatro tipos de características diferentes e avaliados os resultados de sistemas construídos a partir de cada uma, e de sua associação.

6.2.1. Energia (ENE)

A energia de um sinal temporal discreto é uma medida do grau de distanciamento dos valores com relação à sua média, e é definida como:

$$ENE = \sum_{i=1}^N x_i^2$$

Sendo $X = \{x_1, x_2, \dots, x_N\}$ = Sinal de tamanho N

ENE = Energia do sinal x

Para cada um dos sinais é obtido um valor de energia total. Teoricamente para o sinal padronizado onde $x_i \in [-1, 1]$ o valor de energia deve cumprir $ENE < N$.

6.2.2. Entropia (ENT)

A entropia é fundamentalmente uma medida da aleatoriedade ou imprevisibilidade nos valores de uma sequência de pontos de uma distribuição discreta, originariamente apresentado por Claude Shannon como parte fundamental da Teoria da Informação. Além de outras áreas, o processamento de sinais faz uso do conceito de entropia como medida da “desorganização de um sistema”, e é também uma medida da dispersão das informações contidas num vetor. A expressão de entropia aplicada a um vetor N-dimensional, como neste trabalho, é definida como segue (ADDISON, 2002):

$$ENT = - \sum_{i=1}^N P_i \log_2 P_i$$

$$P_i = \frac{x_i^2}{\sum_{n=1}^N x_n^2}$$

Sendo $X = \{x_1, x_2 \dots x_N\}$ = Sinal de tamanho N

ENT = entropia do sinal X

P_i = energia normalizada do componente x_i do sinal

Com a expressão anterior, e considerando cada um dos sinais exemplo como o vetor X, é calculado um valor de entropia para cada sinal.

Se a grandeza dos valores do sinal seguisse uma distribuição uniforme o valor da entropia seria determinado somente pelo tamanho do vetor. O valor mínimo da entropia se apresentaria no caso hipotético de somente um elemento do vetor ser diferente de 0. Assim, e lembrando que os sinais adquiridos são padronizados de forma a ter valores $x_i \in [-1,1]$, o valor da entropia calculado será necessariamente $ENT < 7.9$.

6.2.3. Número de cruzamentos por zero (NCZ)

O número de cruzamentos por zero, como o próprio nome indica, nada mais é que a contagem do número de ocorrências de um elemento positivo aparecer imediatamente depois de um elemento negativo, ou um negativo depois de um positivo.

$$NCZ = \sum_{i=2}^N I(x_i \cdot x_{i-1} < 0)$$

Sendo $X = \{x_1, x_2 \dots x_N\}$ = Sinal de tamanho N

NCZ = Número de cruzamentos por zero do sinal X

$I(s) = 1$ se s = Verdadeiro

$I(s) = 0$ se s = Falso

Lembrando que os sinais são padronizados, a característica de número de cruzamentos por zero é uma medida da frequência de oscilações ao redor da média. Comumente o ruído no sinal aparece como oscilações de alta frequência gerando, portanto, valores altos no número de cruzamentos por zero. Assim, quanto maior seja a influência do ruído no comportamento do sinal, maiores serão os valores de NCZ. O resultado para cada sinal é uma quantidade discreta entre 1 e N-1.

6.2.4. Distribuição da energia na decomposição *wavelet* (EDW)

Na seção 4.2 foram apresentados os conceitos básicos relativos à transformada *wavelet*. Ao executar a decomposição *wavelet* discreta, o sinal original é representado por dois novos componentes ou vetores de coeficientes ordenados no tempo, e correspondentes às frequências baixas (aproximação) ou a frequências altas (detalhes), respectivamente. O processo continua iterativamente até o nível escolhido, dividindo sucessivamente a componente de aproximação do nível anterior em duas novas componentes. Como resultado, decompondo até o nível m o sinal fica representado por um vetor de coeficientes da aproximação, e m vetores de coeficientes dos detalhes, completando em total $m+1$ vetores ou componentes.

Para analisar os resultados da decomposição *wavelet* estuda-se em qual das componentes está concentrada mais ou menos informação, o que pode ser medido com os valores da energia definida anteriormente (6.2.1). É calculado o valor da energia correspondente a cada um dos componentes resultantes da decomposição, e expresso como percentagem da energia total dos $m+1$ componentes. A característica de distribuição da energia nas componentes da decomposição *wavelet* (EDW) do sinal corresponde ao vetor de valores resultante, composto de $m+1$ números reais de 0 até 100.

6.3.CLASSIFICADOR

O classificador é a componente do sistema de reconhecimento de padrões através do qual são analisadas as características extraídas e é definida a resposta, que neste trabalho é a classe (COM ou SEM) a que o sinal apresentado pertence. A construção do classificador (chamada de treinamento em alguns casos) consiste da definição dos padrões das classes analisadas, o método de comparação do novo sinal com eles e a designação da classe. Os classificadores considerados estão baseados em aprendizado supervisionado, portanto sua construção é baseada em exemplos classificados, isto é, o conjunto de sinais da classe COM e o da classe SEM. O objetivo do classificador é alocar todo sinal novo à classe adequada, mas é realmente difícil que com o algoritmo consiga-se definir exatamente essa resposta, e em geral é avaliada a probabilidade do sinal pertencer a cada uma das classes para depois relacioná-lo à classe mais provável (DUDA; HART; STORK, 2001).

A representação dos sinais pela extração de características permite a aplicação de grande quantidade de metodologias de classificação cuja conceituação e desenvolvimento independem da natureza dos sinais. Com o objetivo de avaliar a utilidade de diferentes classificadores dentro do sistema de reconhecimento de padrões para detecção de vazamentos, vários tipos de metodologias foram aplicados. A seguir são apresentados os conceitos básicos de cada um desses classificadores, e os resultados da sua implementação são avaliados segundo indicado em 6.4 e discutidos em 7.4.

6.3.1. Vizinhos mais próximos

O algoritmo dos k vizinhos mais próximos (KNN do inglês *K Nearest Neighbour*) é considerado o mais simples de todos os algoritmos de aprendizado de máquina (FACELI et al., 2011). É um classificador linear de uso expandido, que é subótimo já que conduz a taxas de erro maiores que o mínimo possível teoricamente (THEODORIDIS; KOUTROUMBAS, 1999). A ideia fundamental do algoritmo é que os elementos de uma mesma classe devem ter valores das características semelhantes em comparação com os elementos de classes diferentes.

Na apresentação do algoritmo KNN e de muitos outros dos classificadores é importante o conceito do espaço de características (feature space) ou espaço de entrada. O espaço de entrada é um espaço V-dimensional definido pelos V atributos extraídos dos sinais. Cada um

dos sinais de exemplo adquiridos, representado pelo seu vetor de características de tamanho V , pode então ser visto como um ponto no espaço de entrada, ou um vetor desde a origem até esse ponto, e que tem associada a sua classe já conhecida (BEALE; JACKSON, 2010). Na Figura 14 é apresentado um exemplo de espaço de entradas usando vetores de características de dimensão três (CAR.1, CAR.2, CAR.3) para elementos pertencentes a duas classes (A e B).

Figura 14. Espaço de entrada para elementos de dimensão três pertencentes a duas classes

O algoritmo KNN é do tipo *lazy*, baseado em memória ou preguiçoso, ao não criar uma representação compacta ou modelo dos objetos, dado que não contempla nenhum processamento anterior à entrada de um novo elemento a ser classificado. Assim, o treinamento do classificador consiste simplesmente em memorizar todos os exemplos. Os passos básicos do algoritmo para classificação são:

- Escolher os k elementos (vetores) mais próximos, quer dizer com menor distância de x_0 , independentemente da classe à que pertencem. A definição do melhor valor de K é abordada na seção 7.2.

A determinação da distância entre o novo elemento (x_0) e os demais elementos (x_j), no espaço de entrada, pode ser obtida usando a distância de Mahalanobis, Minkowski, Manhattan, Chebyshev ou, muito mais comumente, a distância Euclidiana definida como segue:

$$d(x_0, x_j) = \sqrt{\sum_{v=1}^V (x_0^v - x_j^v)^2}$$

Sendo $d(x_0, x_j)$ = Distância entre os vetores x_0 e x_j

x_0, x_j = Vetor de entrada y um dos de exemplo, respectivamente. Tamanho V

x_0^v, x_j^v = valor dos vetores x_0 e x_j na coordenada v

- Dentre os k elementos de exemplo escolhidos, identificar quantos deles correspondem a cada uma das classes.
- Alocar o elemento novo x_0 à classe correspondente à maior quantidade dos k exemplos, ou à moda das classes dos vizinhos mais próximos. Variações do algoritmo podem incorporar uma ponderação para a classe de cada k exemplo dependente da sua distância.

Segundo o algoritmo KNN, como todos os classificadores baseados em distância, a escala dos valores é um fator muito importante. Por exemplo, dentre as características apresentadas na seção 6.2 algumas estão na faixa de 1 até 100 enquanto outras são sempre menores do que 10. Considerando que uma distância proporcionalmente pequena na coordenada WDE teria uma importância muito maior que uma distância proporcionalmente alta na coordenada ENT, para evitar essa assimetria, os valores das características são padronizados antes de executar o algoritmo de classificação. A padronização é realizada para que todas as características fiquem expressas no intervalo de 0 até 1, segundo:

$$x_i = \frac{y_i - a_i}{b_i - a_i}$$

Sendo $y = \{y_1, y_2 \dots y_M\}$ = vetor de características sem padronizar

$x = \{x_1, x_2 \dots x_M\}$ = vetor de características padronizado

a_i, b_i = limites inferior e superior, respectivamente, dos valores possíveis para a i-ésima característica no vetor y, para $i = 1, 2 \dots M$

6.3.2. Árvore de decisão

As árvores de decisão são sistemas de classificação (ou de regressão em outros casos) baseados na avaliação sequencial dos valores das características de um novo elemento, até

chegar finalmente a uma classe na qual o elemento possa ser alocado. A definição das regras sequenciais que constituem uma árvore de decisão corresponde com a divisão sucessiva do espaço de entrada (seção 6.3.1) em regiões não intersectadas menores que correspondem às diferentes classes. Essa divisão é feita normalmente com hiperplanos de lados paralelos aos eixos, isto quer dizer que cada uma das regras para a divisão corresponde a uma pergunta binária relativa a uma característica, do tipo “é a característica C menor que o valor Y?”. Existem também outros tipos de funções mais complexas, mas sua aplicação prática é menos comum.

Uma árvore de decisão é constituída por um conjunto de nós ordenados, a partir de um único nó de decisão, que desempenham uma sequência de testes. Cada nó de decisão está unido a outros dois nós, relativos às respostas para o teste, e as uniões com eles são excludentes, pois só uma delas pode ser seguida. Os nós posteriores podem ser também de decisão com a mesma estrutura anterior, ou nós de folha. Os nós de folha correspondem diretamente a uma classe. Um esquema de uma árvore de decisão para as duas classes COM e SEM com um vetor de características de tamanho quatro é apresentado na Figura 15. A árvore pode ser representada também como um conjunto de regras, cada regra começando no mesmo nó inicial, mas continuando por diferentes percursos até uma folha diferente.

Figura 15. Exemplo de árvore de decisão para duas classes, com quatro características.

O processo de treinamento do classificador consiste na definição dos nós e sua sequência baseada no conjunto de exemplos classificados, o que equivale à divisão do espaço de entrada. A classificação de um novo elemento consiste na busca sequencial da região no

espaço a que ele corresponde. Para tanto, seus valores são submetidos à sequência de testes da árvore de decisão até chegar a um nó folha.

As árvores de decisão têm como vantagem fornecer um modelo inteligível do processo de classificação, que auxilia na definição das características mais importantes para a distinção entre classes ou ainda na avaliação da definição das classes. Adicionalmente é possível analisar o processo de classificação de um elemento ou instância particular, acompanhando seu percurso pelos nós da árvore até um nó de folha. Seu uso é mais comum em problemas com várias classes, e para problemas com poucas características torna-se menos interessante. Existem diversos algoritmos para a definição de árvores de decisão, no trabalho foi utilizado o C4.5 (QUINLAN, 1993), segundo o qual o ordenamento das regras é feito considerando a entropia.

6.3.3. Regras de Decisão

Analogamente ao descrito para as árvores de decisão, os classificadores baseados em regras também podem ser interpretados como a divisão do espaço de entrada em regiões, cada uma referida a uma única classe. Mas diferentemente das árvores, as regras de decisão não correspondem com um processo de recursivo de divisão, e sim uma divisão direta. Assim, cada regra cobre uma região ou subconjunto de exemplos, e para classificar um novo elemento todas as regras são avaliadas para definir em qual a região e, portanto, a classe à qual pertence. As regras podem ser ordenadas ou não; no primeiro caso, (ordenadas) para cada elemento novo, as regras são avaliadas numa ordem pré-definida até que alguma delas seja cumprida. No segundo caso (não ordenadas), todas as regras são avaliadas e coletadas as que são cumpridas; se elas se referem à mesma classe, a classificação chega ao fim, mas se foram cumpridas regras relativas a diferentes classes um processo adicional de avaliação da classe mais provável deve ser realizado.

Uma variação interessante da classificação com regras de decisão é o algoritmo de uma única regra (1R ou OneR, do inglês *One Rule*), apresentado por Holte (1993). Através da experimentação com diversas bases de dados foi comprovado que o mesmo desempenho dos classificadores comuns, como as árvores de decisão, pode ser atingido também com um algoritmo muito mais simples baseado numa única regra, que pode ser entendida como uma árvore de decisão para um único atributo. Para cada qual dos atributos, a faixa total de valores é dividida em uma quantidade finita de intervalos dentro dos quais estejam contidos pelo

menos um número pré-determinado de elementos da mesma classe (6 elementos, comumente) e o mínimo de elementos de outras classes. O algoritmo cria então um conjunto de regras relativo a cada atributo independentemente, segundo os intervalos definidos, e cada uma das regras é avaliada e calculada a taxa de erro correspondente. O conjunto de regras do atributo que apresentar o menor erro será o único utilizado, descartando totalmente os outros. Se usado um vetor de características de um único elemento, o conceito básico é igual para o 1R e a árvore de decisão, sendo diferente só o algoritmo usado para a definição das regras, que gera resultados diferentes.

6.3.4. Naive Bayes

Uma aproximação interessante para o problema da classificação é a baseada na natureza estatística das características extraídas, colocando o problema em termos de probabilidade e supondo que parâmetros das distribuições que regem essas probabilidades podem ser conhecidos. O principal classificador probabilístico é o Naive-Bayes, baseado na teorema de Bayes de decisão.

A idéia central do classificador probabilístico é decidir a classe que tenha a maior probabilidade de representar um novo sinal analisado, usando os padrões estabelecidos para cada classe predefinida através da análise de exemplos classificados. Estas técnicas estão baseadas na quantificação do melhor resultado entre várias possíveis decisões de classificação, usando a probabilidade e os custos associados a cada decisão.

A regra de decisão de Bayes para a classificação de um elemento (no caso, um vetor de características) é definida a partir da minimização da probabilidade de erro na classificação, e pode ser definida como segue.

Dado x , a classe escolhida para ele é:

$$w_i \text{ se } P(w_i|x) > P(w_j|x) \text{ para todo } j \neq i$$

Sendo: x = o elemento a ser classificado (vetor de características)

w = conjunto de classes, definido com antecedência, de tamanho N

$P(w_j|x)$ = Probabilidade *a posteriori*, ou probabilidade da classe w_j dado x

A probabilidade *a posteriori* está definida pela expressão a seguir (fórmula de Bayes):

$$P(w_j|x) = \frac{P(w_j)p(x|w_j)}{p(x)}$$

Sendo: $P(w_j)$ = Probabilidade *a priori* da classe w_j

$p(x|w_j)$ = Probabilidade condicional. Expressa a função de densidade de probabilidade de x dada a classe w_j

$p(x)$ = Função de densidade de probabilidade de x . Pode ser visto como um fator de escala que garante que a soma das probabilidades posteriores seja 1. É definida como:

$$p(x) = \sum_{j=1}^N p(x|w_j)P(w_j)$$

A probabilidade a priori de cada classe deve ser definida para o problema ao qual se aplica, por exemplo, segundo a frequência relativa de cada classe no conjunto de exemplos, assumindo distribuição uniforme das classes ou adotando valores diferentes segundo conhecimento adquirido sobre a ocorrência das classes no fenômeno real. Nas análises seguintes foi utilizada a probabilidade a priori de $P(COM) = P(SEM) = 0.5$

Para o cálculo da probabilidade condicional é necessário definir uma função de densidade de probabilidade que ajuste os exemplos. Rotineiramente é assumida a distribuição normal com média e desvio padrão calculados a partir do conjunto de exemplos de cada classe, $P(x|w_j) = N(\bar{x}, \sigma_x)$. Outros tipos de distribuição também podem ser usados na procura de um melhor ajuste aos exemplos.

Como a função de densidade de probabilidade $p(x)$ é igual para todas as classes, e no classificador a decisão é tomada segundo a comparação de probabilidades a posteriori de diferentes classes e não segundo seus valores absolutos, o cálculo de $p(x)$ pode ser negligenciado.

6.3.5. Máquina de vetores de suporte

As máquinas de vetores de suporte (SVM, do inglês *Support Vector Machine*) são sistemas de aprendizado que podem realizar tarefas de classificação, usam um espaço de atributos de grande dimensão, e são treinadas com um algoritmo de aprendizado baseado em otimização e na teoria de aprendizado estatístico (TAE) (FACELI et al., 2011). Para alguns autores as SVM são casos particulares de Redes Neurais Artificiais (REZENDE, 2003).

De maneira geral, as SVM transformam o conjunto de treinamento do espaço de entradas (cuja dimensionalidade é igual ao tamanho do vetor de características usado) para um novo espaço de maior dimensão denominado espaço de atributos (que não é mais gerado pelas características extraídas como é o espaço de entrada). Uma vez no espaço transformado as classes devem tornar-se linearmente separáveis, sendo possível definir um hiperplano que divida as classes. Esses hiperplanos de divisão são achados usando uma SVM linear, salientando que a função obtida será linear só no espaço transformado, mas não necessariamente linear no espaço de entrada.

As funções de mapeamento ou transformação para o espaço de dimensão maior podem ser lineares (para problemas com classes linearmente independentes no espaço de entrada), polinomiais, gaussianas, radiais ou outras, e sua escolha deve ser analisada para cada problema. Entretanto, a dimensionalidade do espaço obtido pode ser arbitrariamente grande, normalmente limitada pelo custo computacional. A escolha dos hiperplanos de separação ótimos no novo espaço é o resultado do treinamento de uma SVM; para esses fins o hiperplano ótimo é aquele que tem a maior distância ou margem em relação aos elementos de exemplos mais próximos. Assim, o treinamento envolve a solução de um problema de otimização quadrático para maximizar essa margem ou distância. Os elementos mais próximos, que conservam a mesma distância (mínima) do hiperplano de separação, são de particular interesse e chamados de vetores de suporte. A definição anterior do hiperplano ótimo conduz à melhor capacidade de generalização da SVM, uma das suas principais vantagens. Porém, a sensibilidade aos valores escolhidos para os parâmetros e a criação de um modelo não interpretável são desvantagens a serem consideradas (DUDA; HART; STORK, 2001). Depois do processo de treinamento, um novo elemento que ingressa à SVM será comparado com os limites dos hiperplanos achados para decidir a classe à qual será atribuído.

6.3.6. Rede Neural Artificial de Funções de Base Radial

O termo Rede Neural Artificial (RNA, ou ANN do inglês *Artificial Neural Network*) refere-se a grande quantidade de modelos matemáticos estruturados na forma de unidades de processamento simples (neurônios) densamente interligados, nos quais o armazenamento e processamento da informação são feitos de maneira paralela e distribuída. Além de muitas outras aplicações, as RNA podem ser utilizadas como classificadores. No modelo, as unidades ou neurônios estão organizados em camadas de processamento sequenciais, sendo que o vetor de características fornece a entrada para a primeira camada e um valor referente à classe é a

saída da última camada. Cada neurônio executa uma função usando um ou mais valores de entrada afetados por pesos que ponderam a entrada recebida, e produz um valor de saída que poderá ser usado como entrada na camada seguinte. O treinamento da RNA como classificador, aplicando aprendizado supervisionado, consiste no seu ajuste para que reproduza com um erro mínimo as saídas conhecidas (classes), dadas as entradas de exemplo (vetores de características). Posteriormente, a RNA deverá prever a saída (classe) de qualquer entrada nova.

As RNA de funções de base radial (FBR) são redes supervisionadas caracterizadas pelo tipo de funções usadas na camada intermediária, sua arquitetura básica é apresentada na Figura 16. Cada neurônio é determinado por uma função não linear da distância euclidiana entre o vetor de entrada e um centro predefinido (vetor de referência), sendo funções localizadas e que apresentam base radial sobre seu domínio (CASTRO; CASTRO, [s.d.]). Por ser de domínio localizado, para cada vetor de entrada só algumas unidades apresentam saídas significativas, o que é ajustado dentro do processo de treinamento. Essas funções podem ter várias formas, sendo a mais comum a gaussiana:

$$e^{-dist(X,W_i) \cdot b}$$

Sendo X = vetor de entrada

W_i = vetor de referência ou centro do neurônio i

b = fator de dispersão da função

A camada posterior realiza uma combinação linear das saídas das funções de base radial, usando outro vetor de pesos, para gerar a saída. As saídas produzidas pela combinação linear podem ter qualquer valor dentro da faixa estabelecida pela função de ativação usada no neurônio de saída, comumente uma função logística. Assim, para um classificador binário uma das classes pode corresponder a -1, a outra classe a 1, e as respostas do classificador para uma nova entrada serão um número intermediário. A classe escolhida em cada caso será aquela cujo valor de saída estabelecido esteja mais próximo do valor calculado pelo classificador para o vetor de entrada.

Figura 16. Esquema de rede de Funções de Base Radial

6.4. AVALIAÇÃO DO PROCESSO

Um “bom” sistema de reconhecimento de padrões, neste trabalho, será aquele que ao receber um novo sinal produzirá a sua classificação corretamente, mas para avaliar quão “bom” é o sistema não é válido usar um juízo apreciativo. Para qualquer análise científica o desempenho deve ser testado e quantificado com metodologias apropriadas. A seguir, as metodologias usadas neste trabalho são brevemente apresentadas.

6.4.1. Validação cruzada (*k-fold cross validation*)

Uma das propriedades importantes de um sistema de reconhecimento de padrões é a sua capacidade de generalização, isto é, a capacidade de fornecer respostas certas para entradas diferentes daquelas usadas no processo de treinamento. Assim, para avaliar o desempenho é necessário fornecer um conjunto de entradas ao sistema para analisar sua resposta em comparação com as respostas certas previamente conhecidas, mas garantindo que essas informações não tenham feito parte do treinamento. Adicionalmente, existe interesse em evitar que a escolha desse conjunto de entradas para teste influa na avaliação.

Assim, uma metodologia amplamente utilizada é a validação cruzada (*k-fold cross validation*) segundo a qual o conjunto total de exemplos é dividido em k partes de aproximadamente igual tamanho, sendo $k-1$ deles destinados ao processo de treinamento e 1 deles à realização de testes. Portanto, o sistema é treinado com $k-1$ partes e testado com a parte restante, produzindo os indicadores de desempenho de interesse. Posteriormente, o processo é repetido escolhendo outra das partes para teste e as restantes $k-1$ para treinamento, e assim segue até ter usado todas as k partes para teste uma vez. O desempenho será então calculado como a média dos k valores dos indicadores obtidos nas iterações. Na divisão do conjunto de exemplos procura-se que cada parte seja formada por exemplos das diferentes classes na mesma proporção que elas se apresentam no conjunto total.

Em todas as análises da seção 7 foi utilizada a metodologia antes descrita para validação cruzada, com 10 partes, cada uma composta por exemplos das classes COM e SEM na mesma quantidade.

6.4.2. Matriz de confusão

Em problemas de classificação, particularmente binários, é muito útil construir a chamada Matriz de Confusão, em que os resultados de todos os testes são classificados considerando qual a classe correta (linhas) e qual a classe prevista pelo sistema (colunas), como a mostrada na Tabela 3. Normalmente ao falar do problema binário refere-se uma classe como positiva e a outra como negativa, no presente trabalho a classe “positiva” é a referente ao vazamento, COM, enquanto a classe negativa é a referente à ausência de vazamento, SEM.

Assim, formam-se quatro categorias: Verdadeiros COM (VC, exemplos COM classificados corretamente), Falsos COM (FC, exemplos SEM classificados como COM), Verdadeiros SEM (VS, exemplos SEM classificados corretamente) e Falsos SEM (exemplos COM classificados errado), sendo que $VC+FC+VS+FS = N$, número total de sinais de teste.

Tabela 3. Esquema da matriz de confusão

Classe Verdadeira		Classe Obtida	
		COM	SEM
	COM	VC	FS
	SEM	FC	VS

Para um sistema de reconhecimento ideal, tanto FS quanto FC devem ser iguais a zero, enquanto VC e VS devem ter valor de $N/2$ (dado que, neste trabalho, na validação cruzada é garantido que cada conjunto de teste tenha a mesma quantidade de exemplos de cada classe). Inicialmente, a matriz de confusão alerta sobre resultados tendenciosos quando há grande assimetria, mas a sua maior utilidade está no cálculo de indicadores de desempenho como os apresentados a seguir (FACELI et al., 2011).

- Taxa de acerto: Indica o nível geral de acurácia do sistema, contabilizando todos os exemplos classificados acertadamente.

$$\frac{VC + VS}{N}$$

- Taxa de erro: Complemento da taxa de acerto, a taxa de erro é uma medida dos erros cometidos nas duas classes.

$$\frac{FC + FS}{N} = 1 - \text{Taxa de Acerto}$$

- Sensibilidade ou Taxa de verdadeiros positivos: Fração dos exemplos que verdadeiramente são da classe COM e que foram classificados acertadamente.

$$\frac{VC}{VC + FS}$$

- Taxa de falsos positivos: Proporção dos exemplos que verdadeiramente são da classe SEM, mas foram classificados como COM.

$$\frac{FC}{VS + FC}$$

- Especificidade: Fração dos exemplos que verdadeiramente são da classe SEM e que foram classificados acertadamente.

$$\frac{VS}{VC + FS}$$

- Precisão: Proporção dos exemplos classificados como COM que verdadeiramente pertencem a essa classe.

$$\frac{VC}{VC + FC}$$

6.4.3. Curva ROC

A curva ROC (*receiver operating characteristic*) é amplamente utilizada como uma representação gráfica do desempenho de classificadores binários. No espaço de curvas ROC (espaço ROC), como apresentado na Figura 17, o eixo horizontal corresponde à taxa de falsos positivos ou FC (relativo à probabilidade de falsos alarmes), enquanto o eixo vertical corresponde à taxa de verdadeiros positivos ou VC (probabilidade de acerto na detecção). Assim, o desempenho de um classificador ao final de um conjunto de testes (isto é, os valores de VC, VS, FC e FS resultantes) é representado como um ponto na curva ROC. Assim, o ponto (0,1) corresponde a um classificador ideal, que obteve classe igual à verdadeira para todos os exemplos. O ponto (0,0) corresponde a um classificador que sempre entrega a classe negativa (SEM), enquanto no ponto (1,1) um classificador produz sempre a classe positiva (COM) como resposta. A diagonal que representa $VC = FC$ corresponde a classificadores que entregam respostas aleatórias.

Figura 17. Exemplo curva ROC

Face ao exposto, ao analisar no espaço ROC o desempenho de dois classificadores, aquele que apresentar menor valor no eixo x e maior no y, será melhor que o outro. Uma forma muito mais robusta de fazer esse tipo de análise é o cálculo da área abaixo da curva ROC correspondente a cada classificador. Na definição de uma curva ROC é considerado que os indicadores de desempenho do classificador mudam ao mudarem os parâmetros para decisão da classe, isto é, a forma como cada algoritmo de classificação processa uma saída (probabilidade calculada no classificador NaiveBayes, ou valor resultante do neurônio de

saída nas RNA-FBR, por exemplo) para prever uma das duas classes possíveis. Cada classificador fica assim representado por uma curva, sendo que para o classificador ideal a curva seria a linha (0,0)(0,1)(1,1) e para o classificador aleatório a linha (0,0)(1,1).

Um indicador de interesse sobre o desempenho de um classificador é a área abaixo da sua curva ROC (AAC). Assim, quanto mais próximo de 1 for o valor, melhor será o desempenho geral do classificador, enquanto um valor de 0.5 na AAC corresponde com o desempenho de um classificador aleatório.

6.5.IMPLEMENTAÇÃO DO SISTEMA DE RECONHECIMENTO DE PADRÕES

O processamento das informações desde sua aquisição junto ao circuito experimental até a obtenção dos resultados contempla várias etapas e para elas, como parte do presente trabalho, foram desenvolvidas e utilizadas diferentes ferramentas computacionais. O processamento foi realizado em blocos, completando cada etapa com todos os sinais de exemplo antes de seguir para a seguinte etapa. Porém, as metodologias utilizadas poderiam ser implementadas sequencialmente para cada sinal, fazendo mudanças nas rotinas.

A aquisição dos sinais foi feita a través do software integrado ao sistema de aquisição (seção 5.2), obtendo as informações em arquivos de extensão “.xls”. O armazenamento e primeira organização da informação foi feita usando planilhas de MSExcel. Posteriormente todos os sinais adquiridos foram ingressados como variáveis no programa *Matlab*³.

Os processos de segmentação, de padronização e extração de características, foram realizados aplicando rotinas previamente desenvolvidas durante este trabalho fazendo uso do Matlab. Dentro das rotinas desenvolvidas foram usadas funções do *Wavelet toolbox*, além das bibliotecas comuns do Matlab. Todos os vectores de características extraídos dos exemplos foram transformados e exportados para uso no software WEKA⁴.

Para o processo de classificação, utilizando os algoritmos de aprendizado implementados no WEKA foram treinados e testados os classificadores para todas as alternativas de arquitectura do sistema de reconhecimento.

³ Do inglês *Matrix Laboratory*, *Matlab* é um ambiente de análise numérica de alto desempenho e linguagem de programação de quarta geração, desenvolvido pelo MathWorks, amplamente utilizado para ensino, pesquisa e aplicações industriais em todo o mundo.

⁴ *Waikato Environment for Knowledge Analysis – WEKA*, é um popular compendio de software para aprendizado de máquina, desenvolvido em linguagem JAVA, pela Universidade de Waikato, Nova Zelândia.

Finalmente, os resultados do sistema de reconhecimento de padrões com todas as combinações analisadas de opções nas três etapas (segmentação e padronização, extração de características com múltiplas alternativas, e classificação com múltiplas alternativas) são exportado novamente ao software MSExcel para sua análise estatística, geração de gráficos e apresentação dos resultados.

7. RESULTADOS

Como apresentado no capítulo anterior, o sistema de reconhecimento de padrões é constituído basicamente por três módulos sequenciais (segmentação e padronização, extração de características, e segmentação), e em cada um deles diferentes metodologias podem ser aplicadas para conseguir conjuntamente o melhor desempenho final. As análises realizadas para decidir as melhores alternativas na construção do sistema de reconhecimento de padrões e as avaliações posteriores são apresentadas neste capítulo.

Na seção 7.1 são analisadas variações no cálculo de uma das características a ser consideradas na etapa de extração de características, enquanto na seção 7.2 são analisados parâmetros para um dos classificadores propostos. A comparação geral e a avaliação da melhor alternativa para a etapa de extração de características é realizada na seção 7.3, e o correspondente à etapa de classificação é analisado na 7.4. Com os resultados anteriores, fica definida a melhor arquitetura de sistema de reconhecimento de padrões proposta, e seu desempenho é avaliado na seção 7.5. Na seção 7.6 é analisada a influência da magnitude do vazamento no desempenho do sistema. Um esquema das análises apresentadas neste capítulo, em relação à sua aplicação no sistema de reconhecimento de padrões, é mostrado na Figura 18.

Figura 18. Esquema das análises apresentadas neste capítulo, em relação às componentes do sistema de reconhecimento de padrões.

Todos os resultados do presente capítulo foram obtidos analisando toda a base de dados de sinais adquiridos junto ao circuito hidráulica experimental como apresentado na seção 5.3, com 620 exemplos em total (310 de cada classe) e para os 18 sensores⁵. Cada resultado apresentado é produto da aplicação do *10-fold cross validation* apresentado na seção 6.4.1. O indicador de desempenho do sistema mais utilizado para as comparações neste capítulo é a área abaixo da curva ROC (AAC), explicada na seção 6.4.3.

7.1. SELECÇÃO DO TIPO E NÍVEL DE DECOMPOSIÇÃO WAVELET PARA A CARACTERÍSTICA EDW

Como apresentado nas seções 4.3 e 6.2.4, a característica EDW é calculada a partir das componentes que representam o sinal depois da aplicação da decomposição *wavelet*, sendo os componentes obtidos até um dado nível. A seleção do melhor nível de decomposição varia de um problema para outro, sendo necessário avaliar conjuntamente o ganho e o custo do cálculo. No caso, o ganho é medido em termos do desempenho do sistema de reconhecimento de padrões quando desenvolvido usando a característica EDW (distribuição da energia na decomposição *wavelet*) calculada até cada nível de decomposição. O custo de cálculo, entretanto, sabe-se que é maior para níveis superiores.

Foram avaliadas as alternativas de uso da decomposição *wavelet* discreta até os níveis 2, 3, 4 e 5, usando em todos os casos a função *wavelet* Daubechies 2 (db2), função escolhida segundo explicado na seção 4.2.2. A avaliação de cada qual dessas quatro alternativas na extração da EDW foi feita treinando e testando o sistema de reconhecimento de padrões completo (segmentação, extração da característica EDW e classificação) para cada um dos sensores, com todos os exemplos adquiridos. Adicionalmente, todos os testes foram repetidos para cada um dos seis classificadores propostos, visando obter conclusões mais gerais e não dependentes dos outros elementos do sistema de reconhecimento. O indicador de desempenho do sistema usado foi a área abaixo da curva ROC (AAC).

Resultados foram obtidos para cada alternativa de nível de decomposição, usando cada um dos seis classificadores propostos na seção 6.3, e para cada um dos 18 sensores. Para consolidar essa grande quantidade de resultados, os valores relativos aos 18 sensores para

⁵ Como antes explicado, todas as análises são conduzidas para os sinais de cada um dos 18 sensores, isoladamente, durante todas as etapas da análise. Somente quando indicado, é apresentada a média dos resultados dos diferentes sensores.

cada alternativa foram resumidos calculando sua média e desvio padrão, que são apresentados na Tabela 4 e na Figura 19.

Tabela 4. AAC média (desvio padrão) para diferentes níveis de decomposição, segundo o classificador.

Classificador	EDW db2 N2	EDW db2 N3	EDW db2 N4	EDW db2 N5
3NN	0.948 (0.025)	0.968 (0.023)	0.97 (0.022)	0.961 (0.021)
Árvore	0.941 (0.031)	0.954 (0.03)	0.951 (0.022)	0.94 (0.032)
Regras	0.898 (0.03)	0.912 (0.034)	0.896 (0.033)	0.881 (0.032)
Bayes	0.963 (0.037)	0.968 (0.05)	0.971 (0.044)	0.968 (0.038)
SVM	0.905 (0.047)	0.932 (0.057)	0.934 (0.056)	0.923 (0.05)
FBR	0.964 (0.041)	0.971 (0.047)	0.975 (0.035)	0.974 (0.028)
Média	0.936 (0.035)	0.951 (0.04)	0.95 (0.035)	0.941 (0.034)

Figura 19. AAC média usando diferentes níveis de decomposição, segundo o classificador.

Lembrando que o aumento no nível de decomposição implica em complexidade e custo computacional maiores, foi calculado o benefício de incrementar esse nível, definindo esse benefício entre níveis consecutivos como o aumento percentual na média da área abaixo da curva ROC, conforme indicado na Tabela 5.

Tabela 5. Aumento percentual na AAC com o aumento no nível de decomposição, segundo o classificador.

Classificador	de N2 a N3	de N3 a N4	de N4 a N5
3NN	1.98%	0.24%	-0.88%
Árvore	1.28%	-0.32%	-1.01%
Regras	1.44%	-1.60%	-1.49%
NaiveBayes	0.48%	0.31%	-0.36%
SVM	2.66%	0.23%	-1.15%
FBR	0.79%	0.40%	-0.17%
Média	1.44%	-0.12%	-0.84%

Para todos os classificadores utilizados, a decomposição até nível três traz um ganho na média da AAC em comparação com a decomposição até nível 2 (de N2 para N3). Esse ganho é maior do que os obtidos ao passar de nível três para nível quatro (de N3 para N4), e maior ainda do que ganho ao mudar de nível quatro para nível cinco. Adicionalmente, só a passagem de nível dois para nível três traz ganhos positivos segundo todos os classificadores usados. Considerando tais resultados, foi definido para o cálculo da característica EDW, usar a decomposição *wavelet* até nível 3.

A decomposição *wavelet packet*, apresentada na seção 4.3, julgou-se potencialmente interessante para o problema em estudo, ao permitir as análises com maior detalhe para toda a faixa de frequências, até as mais altas. Sendo que essa maior quantidade de informações implica num alto custo computacional, é necessário avaliar o benefício de sua aplicação neste trabalho específico. Para tanto, usando a base de dados de 620 exemplos das classes SEM e COM para cada qual dos 18 sensores, foi calculada a decomposição *wavelet packet* até nível 3 usando a função Daubechies 2 (db2). Para cada exemplo, foi calculada a característica EDW, representada por um vetor de 8 elementos cuja soma é 100.

Para avaliar o desempenho da EDW extraída da decomposição *wavelet packet* foi treinado e testado o sistema de reconhecimento de padrões completo, para cada um dos 18 sensores, usando alternativamente cada um dos seis classificadores. Foi calculada a AAC para cada caso, e a média e o desvio padrão dos resultados para os 18 sensores são apresentadas na Tabela 6. Na Figura 20 os resultados são comparados com aqueles obtidos extraindo a característica EDW da decomposição *wavelet* discreta (db2 até nível 3, discreta).

Tabela 6. AAC média (desvio padrão) usando decomposição *wavelet packet*, segundo o classificador.

Classificador	EDW Wavelet Package
3NN	0.972 (0.029)
Árvore	0.946 (0.029)
Regras	0.901 (0.036)
Bayes	0.959 (0.042)
SVM	0.938 (0.06)
FBR	0.972 (0.043)
Média	0.948 (0.04)

Figura 20. Comparação da AAC média usando decomposição *wavelet discreta* e *wavelet packet*.

Os resultados correspondentes a três dos classificadores utilizados (3 vizinhos mais próximos 3NN, Máquina de vetor suporte SVM e rede neural artificial com funções de base radial FBR) mostram que o desempenho do sistema de reconhecimento de padrões com a característica EDW é superior para a decomposição *wavelet packet* se comparada com a decomposição *wavelet discreta*. Para os três classificadores restantes (Árvore de decisão, Regra de decisão e Naive Bayes) o desempenho é superior usando *wavelet discreta* do que usando *wavelet packet*.

Os resultados podem ser melhor entendidos se analisados os vetores de características gerados ao extrair a característica EDW usando decomposição *wavelet packet*. Os valores de percentagem de energia em cada uma das oito componentes têm uma tendência geral que

pode ser mostrada através dos histogramas de cada um dos valores das características, calculados para os exemplos adquiridos com cada sensor. A comparação dos histogramas construídos para cada componente dos respectivos exemplos SEM e COM, é possível através do exame da Figura 21.

Figura 21. Histograma de frequências da distribuição da energia na decomposição *wavelet packet*, para exemplos SEM e COM.

Várias conclusões podem ser extraídas da comparação dos histogramas. A maior parte dos exemplos COM tem a maior percentagem de energia na componente de aproximação, A3, e as percentagens para os componentes de detalhe, D1 a D7, são usualmente menores. Para os exemplos SEM, a percentagem de energia na componente A3 é usualmente menor do que as percentagens nas componentes D1 a D7. Adicionalmente, as distribuições nos componentes de detalhe são mais concentradas para os exemplos COM do que para os exemplos SEM, e

mantém também uma distribuição semelhante nas componentes D1 a D7, com a maior quantidade de casos correspondendo a percentagens de energias baixas. Já para os sinais SEM, parece que a energia está mais concentrada nos diferentes componentes de detalhe, com pouca participação da componente A3. Assim, nos exemplos COM a maior parte da energia do sinal é devida à mudança produzida pelo início do vazamento, melhor representada na componente de aproximação (ou de baixa frequência). Entretanto, nos exemplos SEM a energia do sinal está mais relacionada com o ruído e as oscilações próprias do sistema físico e de aquisição de dados, razão pela qual apresentam maior dispersão e a componente de aproximação torna-se menos importante.

Em adição à análise dos histogramas, o estudo dos primeiros níveis na árvore de decisão reforçam as conclusões. A componente com maior potencial para diferenciar os exemplos das duas classes é a A3. Adicionalmente, as componentes D2 e D3 também permitem uma diferenciação interessante entre ambas as classes.

Em concordância com as definições das decomposições *wavelet* discreta e *packet*, os coeficientes de aproximação (A3) e os de detalhe da menor frequência (D1) são idênticos se aplicarmos as duas decomposições ao mesmo sinal. Os coeficientes D2 e D3 da decomposição *wavelet packet* correspondem à decomposição dos coeficientes D2 da *wavelet* discreta, e D4, D5, D6 e D7 da *packet* com a decomposição do D3 da discreta. Assim, vê-se que a componente mais importante para distinguir as duas classes (A3) é idêntica na decomposição *wavelet* discreta e na decomposição *wavelet packet*, e que as outras componentes de maior interesse (D2 e D3 na decomposição *packet*) têm também uma correspondência clara na decomposição discreta (D2). Como antes dito, um maior detalhamento nas frequências mais altas não se mostra interessante por corresponder o fenómeno estudado a mudanças nas baixas frequências enquanto as altas frequências estão mais influenciadas pelo ruído.

Adicionalmente, os resultados do sistema de reconhecimento de padrões com a característica EDW com decomposição *wavelet packet* apresenta resultados globais piores (com três dos classificadores avaliados) ou levemente melhores (com os outros três classificadores), como analisado anteriormente. Portanto, julga-se desnecessário utilizar a decomposição *wavelet packet* ao invés da decomposição *wavelet* discreta, considerando que a primeira tem maior custo computacional e complexidade, mas as informações adicionais que fornece não são as mais interessantes para os objetivos do sistema de reconhecimento nem se traduzem em melhores resultados globais.

Considerando o discutido nesta seção, nas análises seguintes a característica EDW será avaliada usando decomposição *wavelet* discreta, da família db2, até nível 3.

7.2. PARÂMETROS PARA O CLASSIFICADOR KNN

Dentro das opções consideradas para a etapa de classificação do sistema de reconhecimento de padrões (terceira etapa, depois da segmentação e a extração de características) está o algoritmo dos Vizinhos mais próximos (KNN). Como apresentado na seção 6.3.1, o mais importante parâmetro a ser decidido na aplicação desse método de classificação é o número de vizinhos a serem inspecionados (K). Não existindo uma regra geral para o valor ótimo desse parâmetro, é necessário seu estudo para cada problema de aplicação.

Com o objetivo de estabelecer o valor mais apropriado para K e avaliar a sensibilidade dessa escolha, foram feitos testes do sistema de reconhecimento de padrões envolvendo seis possíveis valores para K (1, 3, 5, 7, 9, 11) e, portanto, seis classificadores. A análise é feita considerando os dados de cada sensor independentemente, usando o conjunto de quatro características (TOD) e também usando somente a característica de distribuição da energia da decomposição *wavelet*⁶ (EDW).

Os resultados obtidos da aplicação do sistema de reconhecimento de padrões para cada um dos sensores, extraíndo o vetor de características TOD, e usando os seis valores de K para o algoritmo KNN, são mostrados na Tabela 7 e na Figura 22. Esses resultados estão em termos da AAC obtida aplicando correlação cruzada com 10 blocos. Uma análise análoga, porém extraíndo somente a característica EDW, fornece os resultados apresentados na Tabela 8 e Figura 23.

⁶Em ambos os casos, usando decomposição *wavelet* discreta até nível 3 com a função db2, como definido na seção anterior.

Tabela 7. AAC usando variações do algoritmo K-NN, depois da extração do vetor de características TOD.

Sensor	1-NN	3-NN	5-NN	7-NN	9-NN	11-NN
P01	0.923	0.968	0.971	0.977	0.976	0.981
P02	0.971	0.993	0.995	0.995	0.997	0.997
P03	0.974	0.993	0.994	0.997	0.997	0.997
P04	0.963	0.984	0.990	0.994	0.993	0.996
P05	0.937	0.979	0.987	0.988	0.988	0.992
P06	0.969	0.989	0.992	0.994	0.995	0.995
P07	0.974	0.986	0.994	0.995	0.995	0.995
P08	0.965	0.982	0.986	0.992	0.993	0.994
P09	0.969	0.987	0.991	0.993	0.992	0.995
P11	0.971	0.988	0.989	0.993	0.995	0.995
P12	0.971	0.989	0.992	0.992	0.992	0.994
P13	0.976	0.990	0.989	0.989	0.993	0.992
P14	0.969	0.984	0.989	0.989	0.989	0.992
P15	0.976	0.987	0.990	0.993	0.994	0.994
P16	0.958	0.982	0.991	0.990	0.992	0.994
Q01	0.906	0.955	0.960	0.968	0.968	0.968
Q02	0.879	0.936	0.952	0.957	0.958	0.956
Q03	0.816	0.898	0.917	0.925	0.931	0.934
Média	0.948	0.976	0.982	0.985	0.985	0.987
Desvio	0.043	0.024	0.020	0.018	0.017	0.017

Figura 22. AAC usando variações do algoritmo K-NN, depois da extração das características TOD.

Tabela 8. AAC usando variações do algoritmo K-NN, depois da extração da característica EDW.

Sensor	1NN	3NN	5NN	7NN	9NN	11NN
P01	0.911	0.959	0.963	0.972	0.976	0.977
P02	0.956	0.986	0.989	0.989	0.992	0.993
P03	0.942	0.983	0.988	0.988	0.990	0.991
P04	0.947	0.982	0.988	0.988	0.987	0.990
P05	0.940	0.972	0.977	0.981	0.986	0.988
P06	0.947	0.981	0.987	0.988	0.989	0.990
P07	0.948	0.974	0.982	0.984	0.987	0.988
P08	0.945	0.970	0.976	0.984	0.986	0.990
P09	0.942	0.976	0.984	0.985	0.989	0.991
P11	0.939	0.979	0.986	0.988	0.988	0.988
P12	0.956	0.983	0.986	0.987	0.987	0.990
P13	0.961	0.980	0.983	0.988	0.989	0.991
P14	0.931	0.977	0.985	0.987	0.987	0.988
P15	0.942	0.972	0.981	0.985	0.988	0.989
P16	0.932	0.972	0.983	0.989	0.989	0.989
Q01	0.856	0.940	0.956	0.970	0.970	0.971
Q02	0.869	0.939	0.948	0.952	0.955	0.954
Q03	0.823	0.892	0.909	0.912	0.918	0.923
Média	0.927	0.968	0.975	0.979	0.981	0.982
Desvio	0.038	0.023	0.020	0.019	0.018	0.018

Figura 23. AAC usando variações do algoritmo K-NN, depois da extração da característica EDW.

Os resultados apresentados permitem concluir que, tanto usando como vetor de características TOD quanto EDW, o desempenho dos classificadores é superior para valores maiores de K, na maioria dos sinais. Como detalhado na Tabela 9⁷, a melhora no desempenho é muito mais significativa para os valores inferiores de K, com um ganho médio de 4,0% ao passar de 1NN para 3NN. Ao estudar valores maiores de K, o aumento de desempenho torna-se menos significativo.

Tabela 9. Incremento percentual na AAC usado variações do algoritmo KNN

Sensor	1NN a 3NN	3NN a 5NN	5NN a 7NN	7NN a 9NN	9NN a 11NN
P01	4.77%	0.44%	0.82%	0.49%	0.05%
P02	2.95%	0.25%	0.01%	0.38%	0.03%
P03	4.09%	0.48%	0.03%	0.20%	0.15%
P04	3.50%	0.61%	-0.03%	-0.05%	0.24%
P05	3.13%	0.51%	0.45%	0.47%	0.17%
P06	3.42%	0.63%	0.08%	0.05%	0.17%
P07	2.60%	0.71%	0.27%	0.31%	0.10%
P08	2.50%	0.60%	0.74%	0.27%	0.34%
P09	3.43%	0.73%	0.18%	0.32%	0.25%
P11	4.00%	0.74%	0.19%	-0.04%	0.01%
P12	2.62%	0.31%	0.10%	0.06%	0.27%
P13	1.90%	0.31%	0.45%	0.12%	0.18%
P14	4.60%	0.82%	0.22%	0.02%	0.07%
P15	2.97%	0.94%	0.44%	0.24%	0.15%
P16	3.96%	1.12%	0.62%	-0.05%	-0.01%
Q01	8.38%	1.62%	1.34%	0.05%	0.11%
Q02	6.93%	0.98%	0.40%	0.22%	-0.06%
Q03	6.96%	1.71%	0.24%	0.62%	0.50%
Média	4.04%	0.75%	0.36%	0.20%	0.15%
Desvio	1.74%	0.41%	0.35%	0.20%	0.14%

Dadas as características do algoritmo KNN, o aumento no valor K (número de vizinhos) multiplica o custo computacional do processo. Na procura de um equilíbrio entre o custo computacional e a melhora no desempenho do classificador, foi determinado que, para o problema tratado e a base de dados usada neste projeto, o valor mais adequado do parâmetro K é 3. Assim, nas análises seguintes o algoritmo KNN usado será o 3NN.

⁷ A Tabela 9 refere-se aos resultados obtidos usando somente a característica EDW, mostrados na Tabela 8 e Figura 23, mas as conclusões são as mesmas se considerados os resultados usando TOD.

7.3.CARACTERÍSTICAS EXTRAÍDAS

Como apresentado na seção 6.2, foram avaliadas quatro características para representar os sinais adquiridos, sendo elas:

- ENE. Energia, um valor representa cada sinal;
- ENT. Entropia, um valor;
- NCZ. Número de cruzamentos por zero, um valor;
- EDW. Distribuição da energia nas componentes da decomposição *wavelet*, vetor de quatro elementos.

Para analisar a tendência das características dos sinais SEM e COM, foram extraídas as características de cada qual dos 620 exemplos das duas classes, para cada um dos 18 sensores. Usando os valores extraídos, foram construídos os histogramas de frequências relativos aos exemplos SEM e COM para cada característica, ou mais exatamente para cada um dos valores do vetor de características (sendo que a característica EDW é representada por quatro valores). Os histogramas foram construídos com os dados de cada sensor isoladamente, e na Figura 24 são apresentados os do sensor mais próximo do local de vazamento (P02) e o do local mais afastado a jusante dele (P16), cuja localização pode ser conferida na Figura 7.

Para obter uma indicação quantitativa do grau de semelhança entre os valores relativos às duas classes, foi calculado o intervalo central de valores de características que contém o 85% dos exemplos, sendo que 7,5% dos exemplos têm valores da característica inferiores aos do intervalo, e o restante 7,5% valores superiores a ele. Os limites desses intervalos são resumidos na Tabela 10 e indicados também na Figura 24.

Tabela 10. Intervalos de valores mais frequentes das características

	P02		P16	
	SEM	COM	SEM	COM
	7.5% - 92.5%	7.5% - 92.5%	7.5% - 92.5%	7.5% - 92.5%
ENE	10.4 - 34.9	2.8 - 76.5	17.3 - 37.8	2.8 - 42.4
ENT	6.2 - 6.8	6 - 7.6	6.7 - 7	6 - 7
NCZ	58.5 - 121.9	1 - 57.5	76.1 - 123.9	44.9 - 92.5
EDW-A3	5.7 - 30.2	23.6 - 92.2	8 - 23.6	25.7 - 65.4
EDW-D3	9.3 - 50	1.3 - 12.6	15.6 - 50.4	9 - 33.8
EDW-D2	18.6 - 37.4	0.8 - 29.4	18.7 - 35.4	7.7 - 25.9
EDW-D1	4.7 - 33.6	0.3 - 34	7.7 - 34.3	1.7 - 30.5

Figura 24. Histogramas dos valores do vetor de características para exemplos SEM e COM.

É possível verificar que para algumas das características, as faixas de valores mais comuns são similares, enquanto para outras características os intervalos estão mais claramente diferenciados. A seguir as observações principais sobre cada característica:

- Para a característica ENE, tanto no sensor mais próximo (P02) quanto no mais afastado (P16) o intervalo de valores mais frequentes para a classe SEM está contido no intervalo da classe COM. Assim, é difícil diferenciar se um sinal pertence a uma ou outra classe com base no seu valor ENE dado que os mesmos valores aparecem frequentemente para as duas classes.
- Quanto à característica ENT, os intervalos de valores para o sensor P02 são muito semelhantes, e para o sensor P16 são quase idênticos. Assim, maior parte de exemplos COM e SEM tem valores de ENT na mesma faixa, razão pela qual essa característica não se mostra interessante para o propósito de classificação.
- O intervalo de valores mais frequentes de NCZ tende a valores menores para os exemplos COM do que para os SEM segundo os dados do sensor P16, e para o sensor P02 constitui conjuntos disjuntos. Considerando, por exemplo, somente os dados do sensor P02, com uma regra simples segundo a qual $NCZ > 58 \rightarrow SEM$ e $NCZ < 58 \rightarrow COM$, 85% dos exemplos seriam corretamente classificados. Esse comportamento é compreensível ao lembrar que a ocorrência do vazamento induz uma mudança no sinal que faz com que os valores se afastem da média (0), apresentando menos cruzamentos por zero em comparação aos exemplos em que essa mudança devida ao vazamento não acontece.
- Quanto à característica EDW, os seus quatro valores apresentam tendências diferentes mas complementares, que não diferem significativamente para os dois sensores. Para a energia A3 o intervalo de valores mais frequentes nos exemplos COM abarca valores maiores do que o intervalo dos exemplos SEM, para os dois sensores considerados. Isso significa que depois da decomposição *wavelet* dos sinais SEM uma parte menor da energia corresponde à aproximação A3 (baixas frequências⁸) e as componentes de detalhes (alta frequência) são mais importantes, sendo esse o comportamento esperado

⁸Vale a pena salientar que quando se mencionam, por exemplo, os “valores mais frequentes” ou os “histogramas de frequências”, a palavra “frequência” refere-se à quantidade de vezes que um valor ou faixa de valores ocorrem, e pode ser quantificada com unidades de número de ocorrências. Por outro lado, quando referidas as “altas / baixas frequências” em referência às componentes da decomposição *wavelet*, trata-se de frequências presentes no sinal, que seria o número de ciclos por unidade de tempo se se tratasse de sinais periódicos, e cuja unidade é a inversa do tempo, 1/sec = Hz.

para um sinal em que a maior parte das oscilações é devida ao ruído, de alta frequência. Entretanto, nos exemplos COM grande parte da energia está concentrada no componente de aproximação A3, devido à mudança provocada pelo vazamento. Os valores de energia nas componentes de detalhe são complementares aos da componente de aproximação ($A3+D3+D2+D1=100$ para cada exemplo individual), portanto para os exemplos SEM os valores de D3, D2 e D1 são maiores do que para os exemplos COM. A diferenciação entre os intervalos de valores mais comuns das duas classes é mais clara para os componentes de menor frequência (D3) do que para os de maior frequência (D1). A importância de cada valor da característica de distribuição da energia na decomposição *wavelet* foi discutida também na seção 7.1.

Os resultados para apenas dois dos sensores foram apresentados acima para melhor compreensão, embora as mesmas análises tenham sido realizadas para todos os outros sensores foram feitas e os seus resultados em nada contradizem as anteriores observações, como pode ser verificado nos histogramas apêndice A.

Os histogramas antes apresentados não fazem parte do sistema de reconhecimento de padrões, e sim de uma análise adicional de comparação entre as características. Para avaliar a utilidade dessas características na representação dos sinais para sua posterior classificação, o sistema de reconhecimento de padrões completo (incluindo segmentação, extração de características e classificação, como descrito na seção 6) foi construído e avaliado, para cada um dos sensores. Na etapa de extração de características seis diferentes alternativas são consideradas, para avaliar as características isoladas, combinadas ou ainda o sinal original, como segue:

- TOD. Todas as quatro características em um vetor de características de dimensão sete;
- ENE. Somente energia, vetor de elemento único;
- ENT. Entropia, vetor de elemento único;
- NCZ. Número de cruzamentos por zero, vetor de elemento único;
- EDW. Distribuição da energia nas componentes da decomposição *wavelet*, vetor de dimensão quatro.
- SIN. Sinal original (sem extrair nenhuma característica), vetor de dimensão 240. O uso do vetor original como vetor de entrada a o classificador é possível somente porque a aquisição em laboratório e o pré-processamento dos exemplos garante que todos eles contenham a mesma quantidade de elementos, sem dados errôneos ou faltantes.

Na etapa de classificação foram usados alternativamente todos os classificadores analisados, visando a descartar variações devidas à sua escolha. Assim, para todos os sensores independentemente, o sistema é construído para cada alternativa de vetor de características com seis classificadores diferentes, definidos na seção 6.3 e que serão avaliados na seção 7.4.

Foi calculada a área abaixo da curva ROC (AAC) para o sistema de reconhecimento de cada combinação considerada, aplicando o procedimento de validação cruzada. Os resultados para todos os sensores e para os diferentes vetores de características considerados são apresentados na Tabela 11, em termos do seu valor médio e desvio padrão para os diferentes classificadores. Nas últimas linhas da mesma tabela apresentam-se valores médios para todos os sensores relativos a cada vetor de características. Os resultados anteriores também são apresentados na Figura 25, na qual cada ponto corresponde à avaliação de uma arquitetura do sistema de reconhecimento de padrões com os dados de um sensor. Cada 18 pontos alinhados verticalmente correspondem aos resultados de todos os 18 sensores segundo um dos seis classificadores (de esquerda para direita: 3NN, Árvore, Regra, Bayes, SVM e FBR). Cada grupo de pontos da mesma cor corresponde à mesma alternativa em termos do vetor de características considerado, isto é, as alternativas a serem comparadas nesta seção para a escolha do melhor vetor de características (TOD, ENE, ENT, NCZ, EDW, SIN).

Figura 25. AAC para diferentes vetores de características segundo o sensor e o classificador considerados.

Tabela 11. AAC média (desvio padrão) para diferentes vetores de características, segundo o sensor considerado.

Sensor	TOD	ENE	ENT	NCZ	EDW	SIN
P01	0.961 (0.02)	0.648 (0.037)	0.6 (0.033)	0.89 (0.05)	0.958 (0.02)	0.831 (0.153)
P02	0.988 (0.01)	0.893 (0.048)	0.877 (0.108)	0.983 (0.012)	0.971 (0.025)	0.838 (0.156)
P03	0.984 (0.014)	0.913 (0.055)	0.908 (0.054)	0.977 (0.016)	0.961 (0.039)	0.838 (0.156)
P04	0.985 (0.013)	0.914 (0.056)	0.915 (0.055)	0.977 (0.019)	0.963 (0.035)	0.841 (0.157)
P05	0.964 (0.025)	0.681 (0.085)	0.641 (0.032)	0.902 (0.042)	0.963 (0.025)	0.839 (0.149)
P06	0.982 (0.016)	0.909 (0.054)	0.911 (0.054)	0.973 (0.023)	0.964 (0.035)	0.841 (0.159)
P07	0.978 (0.023)	0.837 (0.059)	0.858 (0.05)	0.964 (0.024)	0.968 (0.026)	0.843 (0.153)
P08	0.974 (0.019)	0.719 (0.094)	0.694 (0.051)	0.92 (0.042)	0.969 (0.02)	0.836 (0.156)
P09	0.978 (0.024)	0.837 (0.059)	0.834 (0.065)	0.958 (0.03)	0.968 (0.03)	0.819 (0.166)
P11	0.977 (0.025)	0.824 (0.062)	0.849 (0.064)	0.961 (0.027)	0.971 (0.024)	0.837 (0.151)
P12	0.983 (0.016)	0.918 (0.057)	0.916 (0.06)	0.972 (0.021)	0.962 (0.039)	0.829 (0.163)
P13	0.982 (0.015)	0.82 (0.062)	0.814 (0.1)	0.938 (0.034)	0.976 (0.016)	0.855 (0.15)
P14	0.972 (0.024)	0.71 (0.086)	0.739 (0.074)	0.91 (0.047)	0.967 (0.023)	0.828 (0.156)
P15	0.979 (0.018)	0.706 (0.088)	0.74 (0.081)	0.915 (0.043)	0.971 (0.019)	0.839 (0.153)
P16	0.969 (0.018)	0.674 (0.064)	0.664 (0.044)	0.89 (0.038)	0.968 (0.019)	0.816 (0.157)
Q01	0.92 (0.044)	0.851 (0.032)	0.867 (0.024)	0.892 (0.03)	0.885 (0.042)	0.79 (0.122)
Q02	0.904 (0.044)	0.804 (0.05)	0.84 (0.03)	0.78 (0.034)	0.876 (0.068)	0.724 (0.139)
Q03	0.876 (0.035)	0.749 (0.057)	0.791 (0.039)	0.857 (0.027)	0.855 (0.029)	0.85 (0.108)
Media	0.964 (0.022)	0.8 (0.061)	0.803 (0.057)	0.925 (0.031)	0.951 (0.03)	0.828 (0.15)

Na Tabela 11 mostra-se que o uso do vetor de características TOD dentro do sistema de reconhecimento de padrões gera os melhores resultados médios para todos os 18 sensores. Entretanto, a EDW apresenta os segundos melhores resultados médios para 11 sensores e NCZ apresenta essa segunda melhor colocação nos 7 sensores restantes. Agrupando os resultados dos 18 sensores o melhor desempenho médio do sistema corresponde ao vetor de TOD, seguido de EDW e NCZ, que também têm os menores desvios médios, confirmando assim essas três opções de vetor de características como as mais úteis ao sistema de reconhecimento de padrões. Também na Figura 25, pode-se identificar uma tendência geral dos resultados: usando as características ENE e ENT os resultados são em média inferiores e apresentam uma dispersão maior, enquanto usando TOD, NCZ e EDW os resultados são em média maiores e menos dispersos.

O valor de desempenho é afetado pela escolha do classificador, e para avaliar sua influência na Tabela 12 são mostrados os valores de desempenho médios para cada alternativa de vetor de característicos, discriminados segundo o tipo de classificador utilizado. Apesar das variações nos valores, o ordenamento das melhores características é semelhante ao antes identificado, com TOD e EDW na primeira e segunda melhor posição dentre as alternativas de extração de características, superados por pequena diferença pelo uso do vetor SIN quando considerados dois dos classificadores.

Tabela 12. AAC média para diferentes vetores de características, segundo o classificador.

Classificador	TOD	ENE	ENT	NCZ	EDW	SIN
3NN	0.976	0.804	0.812	0.948	0.968	0.984
Árvore	0.963	0.811	0.812	0.902	0.954	0.898
Regra	0.931	0.787	0.796	0.899	0.912	0.766
Bayes	0.979	0.840	0.828	0.953	0.968	0.593
SVM	0.953	0.700	0.715	0.894	0.932	0.942
FBR	0.984	0.859	0.857	0.956	0.971	0.784
Média	0.964	0.800	0.803	0.925	0.951	0.828

Tanto a análise dos histogramas de valores das características quanto os indicadores de desempenho médio e a dispersão dos resultados conduzem às mesmas conclusões a respeito da escolha das características, indicando que os melhores resultados médios são os apresentados pelo uso dos vetores TOD, EDW e NCZ. Mas o vetor TOD está constituído por todas as quatro características avaliadas, incluindo as que se julgam menos interessantes para o reconhecimento (ENT e ENE). Para avaliar o efeito de descartar totalmente essas duas características de piores resultados, foi construída outra alternativa de vetor de características constituída pela combinação de NCZ e EDW. A comparação desse novo vetor de características com os anteriormente identificados como de melhor desempenho é apresentada na Tabela 13.

O uso das características de NCZ+EDW (vetor de 5 elementos) fornece resultados médios muito semelhantes aos do uso do vetor TOD (de 7 elementos), sendo a AAC média de 0.961 no primeiro e 0.964 no segundo. Essa constatação sugere que o cálculo e inclusão das características ENE e ENT não aporta benefícios ao processo de classificação e somente eleva o custo computacional do processo.

Por outro lado, a característica que isoladamente fornece os melhores resultados médios é EDW, mas a incorporação da segunda melhor característica, NCZ, incrementa um pouco os

resultados médios (de 0.951 a 0.961 em média) e diminui levemente os desvios (de 0.030 a 0.023, em média). Como seu cálculo é pouco complexo se comparado com o da EDW, considera-se interessante o uso das duas características combinadas, NCZ+EDW.

Tabela 13. AAC média (desvio padrão) segundo o sensor considerado, para os vetores de características de melhor desempenho.

Sensor	TOD	NCZ	EDW	NCZ+EDW
P01	0.961 (0.02)	0.89 (0.05)	0.958 (0.02)	0.958 (0.022)
P02	0.988 (0.01)	0.983 (0.012)	0.971 (0.025)	0.986 (0.01)
P03	0.984 (0.014)	0.977 (0.016)	0.961 (0.039)	0.98 (0.013)
P04	0.985 (0.013)	0.977 (0.019)	0.963 (0.035)	0.984 (0.012)
P05	0.964 (0.025)	0.902 (0.042)	0.963 (0.025)	0.965 (0.024)
P06	0.982 (0.016)	0.973 (0.023)	0.964 (0.035)	0.975 (0.016)
P07	0.978 (0.023)	0.964 (0.024)	0.968 (0.026)	0.975 (0.023)
P08	0.974 (0.019)	0.92 (0.042)	0.969 (0.02)	0.973 (0.018)
P09	0.978 (0.024)	0.958 (0.03)	0.968 (0.03)	0.974 (0.023)
P11	0.977 (0.025)	0.961 (0.027)	0.971 (0.024)	0.977 (0.023)
P12	0.983 (0.016)	0.972 (0.021)	0.962 (0.039)	0.979 (0.015)
P13	0.982 (0.015)	0.938 (0.034)	0.976 (0.016)	0.978 (0.016)
P14	0.972 (0.024)	0.91 (0.047)	0.967 (0.023)	0.968 (0.023)
P15	0.979 (0.018)	0.915 (0.043)	0.971 (0.019)	0.973 (0.02)
P16	0.969 (0.018)	0.89 (0.038)	0.968 (0.019)	0.97 (0.017)
Q01	0.92 (0.044)	0.892 (0.03)	0.885 (0.042)	0.905 (0.04)
Q02	0.904 (0.044)	0.78 (0.034)	0.876 (0.068)	0.901 (0.073)
Q03	0.876 (0.035)	0.857 (0.027)	0.855 (0.029)	0.867 (0.032)
Media	0.964 (0.022)	0.925 (0.031)	0.951 (0.03)	0.961 (0.023)

Quanto ao uso do sinal original (sem extrair características), SIN, como vetor de entrada para o processo de classificação, cada exemplo é considerado um ponto dentro de um espaço de dimensão 240, e para a definição de distâncias, regras ou distribuições nos procedimentos de treinamento dos classificadores o i -ésimo valor da série é comparado somente com o i -ésimo das outras séries, sendo $i = 1, 2, \dots, 240$. Esse tipo de comparação ponto a ponto tem desvantagens considerando que nos exemplos COM o vazamento acontece em diferentes momentos dentro do intervalo de exemplo. O desempenho do sistema de reconhecimento de padrões usando o vetor SIN foi muito variável dependendo do classificador utilizado, com

valores individuais de AAC desde 0.567 (sensor P16, classificador Bayesiano) até 0.997 (sensor P11, classificador 3NN). Assim, os valores de desvio relativos a SIN com cada sensor mostrados na Tabela 11, de média 0.15 e mínimo 0.11, são maiores do que os desvios de todas as outras alternativas de vetores de características, restando portanto confiabilidade aos seus valores médios. O tempo de processamento usando SIN foi muito maior do que com os outros vetores de características, devido ao grande tamanho dos vetores envolvidos, embora com essa alternativa seja economizado o tempo que leva a extração de características. Como apontado anteriormente, o uso do sinal sem extrair características somente é possível pelo fato de todos os exemplos terem exatamente o mesmo tamanho. Se alguns dos exemplos tivessem tamanho diferente, essa opção de uso do vetor SIN não seria aplicável, enquanto o processo de extração de características não encontraria nenhum problema.

Assim, considerando que o uso do sinal diretamente como se fosse o vetor de características não diminui o custo computacional, têm aplicação limitada, e gera desempenhos de menor precisão e, em média, de menor acurácia, conclui-se que o uso da extração de características é muito mais interessante do que o uso do sinal de entrada, SIN, diretamente.

7.4.TIPOS DE CLASSIFICADOR

Os seis diferentes tipos de classificador apresentados na seção 6.3 foram avaliados como alternativas dentro do sistema de reconhecimento de padrões criado, sendo eles:

- 3NN. Algoritmo de 3 vizinhos mais próximos, padronizando as entradas, usando distância Euclidiana e desconsiderando ponderação;
- Árvore. Árvore de decisão segundo o algoritmo C4.5, com o mínimo de duas instâncias nos nós folha;
- Regra. Algoritmo de uma regra (One Rule), com mínimo seis elementos nos segmentos;
- Bayes. Classificador Naive Bayes, usando distribuição normal;
- SVM. Máquina de vetor suporte, com funções kernel lineares e sem incorporar pesos diferenciados às duas classes;
- FBR. Rede neural artificial com funções de base radial, com funções normalizadas Gaussianas.

Levando em consideração as conclusões da seção 7.3 sobre as características com maior potencial para o processo de reconhecimento de padrões do problema em estudo, nesta etapa é considerado somente o tipo de vetor de características com melhor desempenho: NCZ+EDW. O desempenho para cada classificador e cada sensor, em termos da AAC, é apresentado na Tabela 14 e na Figura 26.

Tabela 14. AAC média (desvio padrão) para diferentes classificadores, segundo o sensor considerado.

Sensor	3NN	Árvore	Regra	Bayes	SVM	FBR
P01	0.967	0.952	0.932	0.985	0.935	0.980
P02	0.990	0.985	0.969	0.993	0.982	0.996
P03	0.985	0.970	0.963	0.991	0.974	0.996
P04	0.987	0.987	0.963	0.992	0.977	0.997
P05	0.974	0.962	0.924	0.988	0.955	0.987
P06	0.981	0.966	0.955	0.991	0.965	0.995
P07	0.983	0.976	0.934	0.994	0.969	0.995
P08	0.974	0.975	0.945	0.993	0.961	0.991
P09	0.981	0.972	0.931	0.994	0.974	0.994
P11	0.984	0.981	0.932	0.995	0.977	0.993
P12	0.987	0.972	0.955	0.991	0.973	0.996
P13	0.982	0.958	0.960	0.995	0.976	0.995
P14	0.979	0.953	0.934	0.993	0.960	0.990
P15	0.981	0.957	0.947	0.994	0.966	0.993
P16	0.971	0.969	0.945	0.991	0.958	0.984
Q01	0.963	0.925	0.865	0.931	0.865	0.884
Q02	0.961	0.962	0.885	0.869	0.777	0.952
Q03	0.906	0.874	0.823	0.859	0.842	0.899
Media	0.974	0.961	0.931	0.974	0.944	0.979
Desvío	0.019	0.026	0.038	0.043	0.057	0.034

Figura 26. AAC para diferentes classificadores, segundo o sensor.

O desempenho de todos os classificadores é muito bom, de forma que as diferenças entre eles são pequenas em relação ao resultado médio. O classificador Bayes apresenta o melhor resultado para nove dos sensores, enquanto o FBR é o melhor para oito dos sensores e o 3NN para os três restantes. Mesmo com diferenças pequenas, os melhores resultados (maior valor de AAC) correspondem aos classificadores FBR, Bayes e 3NN. O desempenho pior foi o da regra de decisão, levemente superado pelo SVM e pela árvore.

Os classificadores de Árvore, Regra e SVM, por terem desempenho médio inferior, são considerados menos interessantes. Mas antes de concluir sobre o classificador mais conveniente, é interessante analisar a aplicação dos diferentes tipos de classificadores desde as suas idéias de base. O classificador Naive Bayes, apresentado na seção 6.3.4, assume que os dados fazem parte de uma distribuição de probabilidades cujos parâmetros são definidos a partir dos dados de treinamento e segundo eles é decidida a resposta dos testes. Adicionalmente, supõe-se independência estatística entre as componentes do espaço de entrada. Comumente a hipótese de independência não se verifica, mas tem sido comprovado em aplicações diversas que isso não prejudica o desempenho do classificador em quanto à

decisão sobre a classe escolhida, embora comprometa o significado dos valores de probabilidade de pertença a cada classe (FACELI et al., 2011).

Quanto ao tipo de distribuição, para os cálculos anteriores foi assumida a distribuição normal, como feito corriqueiramente. Mas ao analisar a verdadeira distribuição dos valores dos vetores de características, como as apresentadas na Figura 24, isso não se verifica. A maioria dos valores das características extraídas não tem comportamento normal, descumprindo a hipótese inicial da aplicação do classificador Naive Bayes. Ainda ao tentar substituir a distribuição normal, nenhum tipo distribuição ajusta bem os dados das diferentes características e sensores simultaneamente. Os resultados em termos do desempenho do sistema de reconhecimento de padrões usando o classificador Naive Bayes podem ser bons devido a que as duas classes têm intervalos de valores mais comuns bem diferenciados, especialmente das características NCZ e EDW (como destacado na seção 7.3), mas o descumprimento da hipótese da distribuição de probabilidades gera desconfiança nestes resultados. Adicionalmente, pode-se prever que com exemplos mais exigentes, para os quais os valores mais comuns das características sejam menos diferentes entre as duas classes, o problema por assumir a priori a distribuição de probabilidade pode ser mais significativo.

Entre os outros dois classificadores de melhor desempenho, o FBR cria um modelo do sistema que não pode ser facilmente interpretado, numa RNA de três camadas, enquanto o 3NN não cria modelo algum, pois usa diretamente a base de dados completa para classificação de cada elemento. Considerando também que o desempenho dos dois classificadores é equivalente, pois a diferença na média é de ordem de grandeza inferior ao desvio, é avaliado que o procedimento mais simples é o mais conveniente. Sendo que o treinamento do FBR implica na adoção de parâmetros que modificam o processo de otimização, com uma complexidade maior que o algoritmo simples do 3NN, sem produzir nem desempenho melhor nem um modelo facilmente interpretável, considera-se mais conveniente usar o classificador 3NN para as análises posteriores.

7.5.RESULTADOS FINAIS DO SISTEMA DE RECONHECIMENTO DE PADRÕES

Com base nas análises relativas aos diferentes componentes do sistema de reconhecimento de padrões e suas alternativas, é possível escolher a arquitetura geral do sistema que fornece os resultados mais interessantes no processo de reconhecimento de sinais adquiridos na ausência

de vazamentos ou durante seu surgimento (classes SEM e COM, respectivamente). Resumindo esses resultados, pode-se definir um sistema de reconhecimento de padrões com as seguintes etapas sequenciais:

- Segmentação e padronização. Segmentação em intervalos de 1 minuto ou 240 pontos, e padronização para valores de -1 a 1, com média 0.
- Extração de características. Extração das características NCZ (número de cruzamentos por zero) e EDW (distribuição da energia nas componentes da decomposição *wavelet* discreta, db2, nível 3) para conformação do vetor de 5 elementos NCZ+EDW.
- Classificador. Classificador 3-NN (três vizinhos mais próximos) padronizando as entradas e usando distância euclidiana sem ponderação.

O sistema assim constituído é avaliado a seguir, considerando as metodologias de avaliação apresentadas na seção 6.4. Lembrando que, como nas seções anteriores, para cada qual dos sinais independentemente o sistema de reconhecimento foi construído e testado usando a metodologia de validação cruzada com dez blocos (*10-fold cross validation*), de forma que para cada sensor têm-se dez testes diferentes. Como nas seções anteriores, os resultados a seguir referem-se às médias dos dez testes realizados como se detalha em cada caso, e os resultados individuais desses testes são apresentados no apêndice B.

7.5.1. Matriz de confusão

Em cada um dos testes, do conjunto de 620 exemplos 10% é usado para testes, sendo 31 exemplos SEM e 31 exemplos COM. A repetição do processo dez vezes, pela validação cruzada, oferece um total de 620 respostas do sistema, com cada sensor. A matriz de confusão dessas 620 respostas é apresentada na Tabela 15 para cada sensor, e também se mostra a matriz média resultante.

Tabela 15. Matrizes de confusão totais para cada sensor, e média.

P01		P02		P03	
285	25	305	5	300	10
15	295	11	299	10	300
P04		P05		P06	
301	9	297	13	297	13
7	303	19	291	11	299
P07		P08		P09	
299	11	295	15	296	14
8	302	14	296	13	297
P11		P12		P13	
299	11	301	9	299	11
8	302	8	302	6	304
P14		P15		P16	
296	14	297	13	300	10
15	295	11	299	15	295
Q01		Q02		Q03	
282	28	281	29	264	46
27	283	24	286	47	263
		MÉDIA			
		294	16		
		15	295		

Tanto as matrizes de confusão dos sensores individuais quanto a média mostram uma simetria razoável com os erros balanceados entre falsos SEM (FS) e falsos COM (FC), descartando viés na classificação. Com base nas matrizes e segundo as definições da seção 6.4.2, os indicadores de desempenho foram calculados para cada um dos sensores e para o conjunto todo, e são apresentados na Tabela 16 e na Tabela 17.

Tabela 16. Taxa de acerto, taxa de erro e sensibilidade do sistema, para cada sensor.

Sensor	Taxa acerto	Taxa erro	Sensibilidade
P01	0.935	0.065	0.919
P02	0.974	0.026	0.984
P03	0.968	0.032	0.968
P04	0.974	0.026	0.971
P05	0.948	0.052	0.958
P06	0.961	0.039	0.958
P07	0.969	0.031	0.965
P08	0.953	0.047	0.952
P09	0.956	0.044	0.955
P11	0.969	0.031	0.965
P12	0.973	0.027	0.971
P13	0.973	0.027	0.965
P14	0.953	0.047	0.955
P15	0.961	0.039	0.958
P16	0.960	0.040	0.968
Q01	0.911	0.089	0.910
Q02	0.915	0.085	0.906
Q03	0.850	0.150	0.852
Média	0.950	0.050	0.949
Desvio	0.031	0.031	0.032

O nível geral de acurácia do sistema, medido pela taxa de acerto, é maior que 0.91 para todos os sensores com exceção do sensor Q03. O valor médio, de 0.95, é satisfatoriamente alto. A taxa de erro é o complemento da taxa de acerto, sendo assim sua média de 0.05. Quanto à sensibilidade ou taxa de verdadeiros positivos, mostra-se que a proporção de exemplos da classe COM classificados acertadamente é superior a 90% para dados de todos os sensores com exceção do sensor Q03. Em média essa porcentagem é de 94,9%, portanto somente 5,1% dos exemplos COM foram classificados errados.

Os falsos positivos correspondem a exemplos SEM erroneamente classificados como COM, e a taxa de ocorrência desses casos é de 0.048, em média. Já a especificidade permite avaliar a capacidade do sistema de identificar os exemplos SEM, calculando a proporção desses exemplos classificados acertadamente, que é de 0.952 em média. A precisão indica que a

porcentagem de exemplos classificados como COM que verdadeiramente pertencem a essa classe é de 95,2%.

Tabela 17. Taxa de falsos positivos, especificidade e precisão do sistema, para cada sensor e média.

Sensor	Taxa falsos positivos	Especificid.	Precisão
P01	0.048	0.952	0.950
P02	0.035	0.965	0.965
P03	0.032	0.968	0.968
P04	0.023	0.977	0.977
P05	0.061	0.939	0.940
P06	0.035	0.965	0.964
P07	0.026	0.974	0.974
P08	0.045	0.955	0.955
P09	0.042	0.958	0.958
P11	0.026	0.974	0.974
P12	0.026	0.974	0.974
P13	0.019	0.981	0.980
P14	0.048	0.952	0.952
P15	0.035	0.965	0.964
P16	0.048	0.952	0.952
Q01	0.087	0.913	0.913
Q02	0.077	0.923	0.921
Q03	0.152	0.848	0.849
Média	0.048	0.952	0.952
Desvio	0.032	0.032	0.032

7.5.2. Área abaixo da curva ROC (AAC)

Para cada um dos sensores pode ser desenhada a respectiva curva ROC, com o eixo horizontal referente a VC (verdadeiros COM) e o eixo vertical a FC (falsos COM). Na Figura 27 são apresentadas, como exemplo, as curvas relativas aos sensores P02, P16 e Q03, enquanto as dos outros sensores são mostradas no apêndice C. O indicador de maior interesse nas análises comparativas neste trabalho é a AAC, cujos valores para cada sensor (na arquitetura definitiva do sistema) são apresentados na Tabela 18.

P02

P16

Q03

Figura 27. Curvas ROC para os sensores P02, P16 e Q03⁹

⁹ Gráficos gerados no software WEKA, como referido na seção 6.5.

Tabela 18. AAC do sistema para cada sensor, e média.

Sensor	Área abaixo da curva ROC
P01	0.967
P02	0.990
P03	0.985
P04	0.987
P05	0.974
P06	0.981
P07	0.983
P08	0.974
P09	0.981
P11	0.984
P12	0.987
P13	0.982
P14	0.979
P15	0.981
P16	0.971
Q01	0.963
Q02	0.961
Q03	0.906
Média	0.974
Desvio	0.019

A AAC (área abaixo da curva ROC), indicador do desempenho do classificador e cujo máximo teórico é um (1.0), apresentou valores altos com todos os sensores, sendo o mínimo 0.906 (sensor Q03), o máximo 0.990 (sensor P02) e média de 0.974, valor satisfatório.

7.6. MAGNITUDE DE VAZAMENTO

Um fator importante para a detecção de vazamentos é a magnitude deles, em relação à vazão na rede. Como apresentado na Tabela 2 da seção 5.3, os experimentos foram conduzidos de maneira a gerar nos exemplos COM diferentes relações entre a vazão do vazamento (V) e a vazão no circuito (Q). Para analisar a influência da relação V/Q no desempenho do sistema de

reconhecimento de padrões criado, foi identificada a faixa de valores V/Q a que pertence cada um dos exemplos usados na etapa de teste dentro do processo de validação cruzada. Analisando os resultados da classificação de cada exemplo para todos os testes, foi calculado o número de exemplos COM classificados acertadamente (VC) ou não (FS) para cada faixa de valores V/Q , como apresentado na Tabela 19. A sensibilidade ou taxa de verdadeiros positivos (definida na seção 6.4.2), é apresentada na Tabela 20. A configuração do sistema de reconhecimento de padrões é a resumida na seção anterior, e inclui segmentação e padronização, extração do vetor de características NCZ+EDW e classificação com 3NN.

A tendência geral que pode ser predita quanto à sensibilidade do sistema de reconhecimento de padrões é que seus valores sejam menores para exemplos com relação V/Q baixa e maiores para os exemplos com relação V/Q maior. Porém, valores individuais em alguns sensores não seguem essa tendência. Por exemplo, a sensibilidade na faixa de 4% a 5% aparece inferior do que a de 3% a 4% para os sinais de vários sensores, sendo o mais crítico o Q01 que aparenta diminuição da sensibilidade em 0.11. Para a faixa de V/Q de 9% a 10% aparece na maioria dos sensores sensibilidade inferior à da faixa anterior (8% a 9%), sendo que essa tendência negativa continua presente ainda nos resultados totais. Uma possível explicação de esse inesperado resultado poderia ser a ocorrência dentro de essa faixa de exemplos extraídos de um experimento com algum erro que tenha gerado corrupção dos dados e dificulte sua classificação, mas essa hipótese não foi comprovada nem levada em conta nas análises executadas com a mesma base de dados de exemplos.

Ao analisar unicamente os resultados de sensibilidade totais, que consideram todos os testes de todos os sensores acumulados, vê-se que, com exceção da faixa de V/Q de 9% a 10%, a tendência esperada é comprovada, sendo que a sensibilidade do sistema de reconhecimento é maior para relações altas de V/Q do que para relações baixas. As diferenças mais importantes estão nas faixas inferiores, notando que o incremento na sensibilidade ao aumentar a relação V/Q é muito mais marcado até as relações de 6%, depois da qual a sensibilidade estabiliza-se. Não pode deixar de ser levado em consideração que, ainda para os exemplos com $V/Q < 3\%$ a sensibilidade é muito alta (0.83) e já com exemplos de V/Q de 4% essa sensibilidade atinge o 0.9. Assim, mesmo sensível à magnitude do vazamento, o desempenho do sistema de reconhecimento de padrões proposto mantém-se razoável mesmo para valores $V/Q < 3\%$.

Tabela 19. Indicadores VC e FS do sistema de reconhecimento calculados para os diferentes valores de V/Q usados nos testes.

Sensor	Indic.	Relação V/Q (%)										Total
		< 3	3 - 4	4 - 5	5 - 6	6 - 7	7 - 8	8 - 9	9 - 10	10 - 11	>11	
P01	VC	2	29	40	37	30	36	29	35	32	15	285
	FS	8	4	11	2	0	0	0	0	0	0	25
P02	VC	8	36	52	35	28	34	33	36	29	14	305
	FS	0	0	1	1	0	0	0	3	0	0	5
P03	VC	10	31	45	36	32	39	29	31	34	15	302
	FS	0	2	3	1	0	0	0	2	0	0	8
P04	VC	10	32	49	36	30	36	29	32	32	15	301
	FS	0	1	2	3	0	0	0	3	0	0	9
P05	VC	7	26	47	38	30	36	29	35	32	15	295
	FS	3	7	4	1	0	0	0	0	0	0	15
P06	VC	9	29	53	38	34	32	31	31	32	12	301
	FS	1	2	2	1	0	0	0	3	0	0	9
P07	VC	10	31	45	40	30	38	23	32	33	13	295
	FS	1	2	5	2	1	0	0	4	0	0	15
P08	VC	8	28	47	39	30	36	29	34	32	15	298
	FS	2	5	4	0	0	0	0	1	0	0	12
P09	VC	6	26	44	39	37	36	29	34	29	16	296
	FS	2	5	6	0	0	0	0	0	1	0	14
P11	VC	9	32	46	38	30	36	29	31	32	15	298
	FS	1	1	5	1	0	0	0	4	0	0	12
P12	VC	11	30	50	37	24	32	34	34	32	17	301
	FS	0	1	3	1	0	0	0	4	0	0	9
P13	VC	9	30	46	38	30	36	29	35	32	15	300
	FS	1	3	5	1	0	0	0	0	0	0	10
P14	VC	9	24	48	41	28	33	26	37	34	16	296
	FS	3	6	3	1	0	0	0	1	0	0	14
P15	VC	10	32	47	38	30	36	29	34	30	14	300
	FS	0	1	4	1	0	0	0	1	2	1	10
P16	VC	10	30	48	37	30	36	29	32	31	14	297
	FS	0	3	3	2	0	0	0	3	1	1	13
Q01	VC	6	32	44	37	29	36	29	35	32	15	295
	FS	4	1	7	2	1	0	0	0	0	0	15
Q02	VC	6	28	46	36	26	31	32	35	28	16	284
	FS	4	4	5	4	1	4	1	2	1	0	26
Q03	VC	7	32	50	34	30	35	25	30	26	14	283
	FS	1	3	3	5	3	0	5	3	3	1	27
TOTAL	VC	147	538	847	674	538	634	523	603	562	266	-
	FS	31	51	76	29	6	4	6	34	8	3	-

Tabela 20. Sensibilidade (taxa de verdadeiros positivos) para os diferentes valores de V/Q usados nos testes.

Sensor	Relação V/Q (%)									
	< 3	3 - 4	4 - 5	5 - 6	6 - 7	7 - 8	8 - 9	9 - 10	10 - 11	>11
P01	0.20	0.88	0.78	0.95	1.00	1.00	1.00	1.00	1.00	1.00
P02	1.00	1.00	0.98	0.97	1.00	1.00	1.00	0.92	1.00	1.00
P03	1.00	0.94	0.94	0.97	1.00	1.00	1.00	0.94	1.00	1.00
P04	1.00	0.97	0.96	0.92	1.00	1.00	1.00	0.91	1.00	1.00
P05	0.70	0.79	0.92	0.97	1.00	1.00	1.00	1.00	1.00	1.00
P06	0.90	0.94	0.96	0.97	1.00	1.00	1.00	0.91	1.00	1.00
P07	0.91	0.94	0.90	0.95	0.97	1.00	1.00	0.89	1.00	1.00
P08	0.80	0.85	0.92	1.00	1.00	1.00	1.00	0.97	1.00	1.00
P09	0.75	0.84	0.88	1.00	1.00	1.00	1.00	1.00	0.97	1.00
P11	0.90	0.97	0.90	0.97	1.00	1.00	1.00	0.89	1.00	1.00
P12	1.00	0.97	0.94	0.97	1.00	1.00	1.00	0.89	1.00	1.00
P13	0.90	0.91	0.90	0.97	1.00	1.00	1.00	1.00	1.00	1.00
P14	0.75	0.80	0.94	0.98	1.00	1.00	1.00	0.97	1.00	1.00
P15	1.00	0.97	0.92	0.97	1.00	1.00	1.00	0.97	0.94	0.93
P16	1.00	0.91	0.94	0.95	1.00	1.00	1.00	0.91	0.97	0.93
Q01	0.60	0.97	0.86	0.95	0.97	1.00	1.00	1.00	1.00	1.00
Q02	0.60	0.88	0.90	0.90	0.96	0.89	0.97	0.95	0.97	1.00
Q03	0.88	0.91	0.94	0.87	0.91	1.00	0.83	0.91	0.90	0.93
Total	0.83	0.91	0.92	0.96	0.99	0.99	0.99	0.95	0.99	0.99

Figura 28. Sensibilidade (taxa de verdadeiros positivos) para diferentes valores de V/Q, total.

8. CONCLUSÕES

Neste trabalho foi desenvolvido um sistema de reconhecimento de padrões para a detecção de vazamentos em rede de abastecimento de água sob pressão pela análise dos sinais, cujas conclusões são aqui apresentadas.

Como conclusão da revisão bibliográfica, foram classificadas as principais técnicas de detecção de vazamentos propostas no âmbito da engenharia hidráulica, e identificou-se que o conjunto de técnicas mais promissor para sua implantação em sistemas reais é o baseado na análise de sinais durante a operação do sistema. A principal vantagem dessas técnicas é não dependerem do conhecimento detalhado do sistema (necessário para o uso de modelos hidráulicos) nem das restritas condições nas quais as análises do transitório são aplicáveis. Porém, seu objetivo é a detecção de novos vazamentos, e não a localização exata nem a identificação de vazamentos existentes. Conclui-se também que apesar dos avanços realizados ainda não existe uma solução final e há muitas ferramentas que podem conduzir a melhores resultados e devem ser pesquisadas, dentre elas o uso de sistemas de reconhecimento de padrões.

Foi adaptado e utilizado o circuito hidráulico experimental, localizado no laboratório de hidráulica ambiental do centro de recursos hídricos e ecologia aplicada (CRHEA). O circuito consiste numa rede de tubulações em policloreto de vinila (PVC) para o escoamento de água sob pressão, e conta com monitoramento automático de pressões e vazões através de 18 sensores ligados a um sistema de aquisição de dados.

Foram executados repetidamente experimentos de duas classes: durante regime permanente sem promover nenhuma mudança no sistema (SEM) e simulando o início de um vazamento com a abertura de um pequeno registro (COM). Os sinais adquiridos constituíram a base de dados usada para todas as análises, com 310 exemplos de cada classe, obtidos com diferentes vazões no circuito (Q) e, no caso dos exemplos COM, diferentes vazões de vazamento (V). Assim, os exemplos COM representam diferentes magnitudes de vazamento, com relação V/Q de 2% até 13%.

O sistema de reconhecimento de padrões é composto por três etapas sequenciais: segmentação e padronização, extração de características e classificação. Depois de conformado o sistema, a entrada é um sinal de pressão ou vazão adquirido e a saída obtida, depois das três etapas antes mencionadas, é a classe a que o exemplo pertence: SEM ou COM. Assim, o sistema permite

determinar, com certo grau de acurácia, se um novo sinal corresponde com uma situação de início de vazamento ou não. No presente trabalho essa avaliação foi feita sempre sobre uma base de dados conformada previamente, mas o sistema desenvolvido poderia ser aplicado à análise de sinais contínuos, em tempo real, precisando para tanto ajustes na rotina computacional.

Para as análises referentes às diferentes componentes do sistema foi aplicado o processo de treinamento e teste do sistema de reconhecimento completo, com os sinais adquiridos nos 18 sensores localizados ao longo do circuito independentemente, produzindo indicadores de desempenho individuais. Adicionalmente, cada avaliação foi feita usando a metodologia de validação cruzada com dez blocos, sendo a média das dez repetições o indicador utilizado em cada caso.

Quanto às primeiras etapas do sistema de reconhecimento de padrões, a segmentação e padronização, os sinais adquiridos foram divididos em segmentos de duração 1 minuto ou 240 valores (taxa de amostragem de 4 amostras/s) e cada segmento constitui um exemplo de uma das classes. Visando a garantir a capacidade de generalização do sistema, as análises foram focadas no comportamento dos sinais e não nas grandezas medidas; assim, cada exemplo foi padronizado de forma que a média fosse 0 e os valores do sinal estivessem entre -1 e 1.

No referente à extração de características, comprovou-se que esta etapa permitiu converter um sinal de grande dimensão a uma representação de dimensão muito menor (vetor de características) e que permite a classificação com acurácia ainda maior que os sinais originais (SIN). As características consideradas foram energia (ENE), entropia (ENT), número de cruzamentos por zero (NCZ) e distribuição da energia nas componentes da decomposição *wavelet* (EDW), sendo as três primeiras calculadas no domínio do tempo e a última baseada na transformada *wavelet*.

Em referência ao cálculo da característica EDW, foi assumido o uso da função Daubechies (db 2), e avaliado o nível de decomposição *wavelet* mais conveniente, considerando que maiores níveis correspondem a maior resolução na frequência, mas maior custo computacional além de menor resolução temporal. Conclui-se que o uso da decomposição até o nível 3, forneceu desempenho médio maior ao fornecido pelo nível 2 e equivalente aos relativos aos níveis 4 e 5. Adicionalmente foi comparado o uso da transformada *wavelet packet* ao invés da transformada *wavelet* discreta, com o intuito de obter maior detalhamento para frequências altas. Concluiu-se que as informações adicionais fornecidas pela transformada *packet* não conduziram a melhorias no desempenho do sistema de

reconhecimento. Foi notado que depois da decomposição *wavelet* (tanto discreta quanto *packet*) a maior energia está concentrada na componente de menor frequência para os exemplos COM, ao contrário dos exemplos SEM, sendo assim essa componente mostrou-se a mais útil na distinção entre exemplos das duas classes. Considerando o anterior, foi concluído que para o cálculo da característica EDW, na etapa de extração de características do sistema de reconhecimento de padrões, fosse usada a decomposição *wavelet* discreta, com a função db2, até nível 3.

Os valores de cada qual das quatro características nas duas classes foram analisadas e foi também avaliado o desempenho do sistema de reconhecimento de padrões com o uso das características isoladas e combinadas. Concluiu-se que as características ENE e ENT foram as menos interessantes para o sistema, enquanto a mais interessante foi EDW, seguida de NCZ. Ao avaliar o uso do vetor de características TOD, conformado pela combinação das quatro características, o desempenho do sistema foi equivalente ao obtido usando o vetor de características da combinação NCZ+EDW; portanto, o aporte das características ENE e ENT ao processo de reconhecimento não é significativo. Foi concluído que na etapa de extração de características fosse calculado o vetor NCZ+EDW como representação de cada sinal.

Para a etapa de classificação foram avaliados seis algoritmos diferentes e foi comparado o desempenho do sistema de reconhecimento de padrões usando cada qual. Antes da comparação, foi analisado o parâmetro K do algoritmo KNN, ou número de vizinhos mais próximos, concluindo pela utilização do valor de $K = 3$. O desempenho do sistema com base nos seis classificadores utilizados (3NN, árvore, Naive Bayes, SVM e FBR) é alto, sendo que as diferenças entre eles são de ordem de grandeza menor à de cada resultado. Os classificadores correspondentes aos melhores desempenhos médios são o FBR (rede neural com funções de base radial), Naive Bayes e 3NN (3 vizinhos mais próximos). Desempenho médio inferior foi obtido usando árvore de decisão, regra de decisão e SVM (Máquina de vetor suporte). Analisando os classificadores de melhor desempenho, foi concluído que os resultados com o Naive Bayes perdem confiabilidade por não serem cumpridas as hipóteses originais quanto à distribuição de probabilidade dos valores das características utilizadas. Por outro lado, tanto FBR quanto 3NN apresentam desempenhos semelhantes, mas nenhum dos dois fornece um modelo interpretável do sistema, razão pela qual a escolha foi feita pela menor complexidade, concluindo o uso do algoritmo 3NN para a etapa de classificação.

Assim, foi concluído que a melhor arquitetura para o sistema de reconhecimento de padrões era: segmentação e padronização, extração do vetor de características NCZ+EDW, e

classificação com 3NN. O sistema foi avaliado com os sinais obtidos de cada sensor individualmente, segundo diferentes indicadores de desempenho produzindo os seguintes resultados médios: taxa de acerto de 95.0%, sensibilidade de 94.9%, taxa de falsos positivos de 4.8%, especificidade de 95.2%, precisão de 95.2% e área abaixo da curva ROC (AAC) de 0.974. Tais resultados indicam um desempenho consistente e de alta acurácia do sistema de reconhecimento de padrões proposto.

Adicionalmente, foi avaliada a influência da magnitude do vazamento sobre processo de detecção. Para tanto, depois de treinar e testar o sistema de reconhecimento de padrões com todos os exemplos, foi calculada a sensibilidade (taxa de verdadeiros positivos) discriminada segundo a relação V/Q dos exemplos de teste. Embora a variação não seja igual para todos os sensores, em média verifica-se que o aumento na relação de V/Q está relacionado com um aumento na sensibilidade. As diferenças mais importantes estão nos exemplos de menor magnitude, com relações V/Q de 3% até 6%, para relações de valores superiores a 6% a sensibilidade não aumentou significativamente. Porém, para relações inferiores a 3%, a sensibilidade média é de 82.6% e para aqueles com $V/Q > 4\%$ a sensibilidade é superior a 90%. Assim, mesmo sensível à magnitude do vazamento, o desempenho do sistema proposto se mantém razoável mesmo para valores $V/Q < 3\%$.

Como mostrado, o sistema de reconhecimento de padrões proposto produziu resultados animadores para o problema de detecção de vazamentos para o qual foi projetado. Por ser este um trabalho de exploração de uma nova metodologia, foram estabelecidos limites ao problema de detecção, desconsiderando, por exemplo, otimização da segmentação para aplicação em tempo real, as variações nas condições do escoamento por razões diferentes ao vazamento ou diferentes localizações do vazamento. Recomenda-se a incorporação de novas condições para comprovar a resposta do sistema proposto, com os ajustes necessários. Considerando os resultados obtidos e a grande quantidade de ferramentas de análise de sinais e aprendizado de máquina, das quais só algumas foram aplicadas neste trabalho, recomenda-se em geral a exploração dessas ferramentas de análise para a detecção de vazamentos em sistemas reais de distribuição de água para abastecimento.

9. REFERÊNCIAS

ADDISON, P. S. The Illustrated Wavelet Transform Handbook. 2002.

AGUIAR, F. G. **Utilização de redes neurais artificiais para detecção de padrões de vazamento em dutos**. São Carlos: Universidade de São Paulo, 2010.

AKSELA, K.; AKSELA, M.; VAHALA, R. Leakage detection in a real distribution network using a SOM. **Urban Water Journal**, v. 6, n. 4, p. 279–289, out. 2009.

ANDERSEN, J. H.; POWELL, R. S. Implicit state-estimation technique for water network monitoring. **Urban Water**, v. 2, n. 2, p. 123–130, jun. 2000.

ASHTON, C. .; HOPE, V. . Environmental valuation and the economic level of leakage. **Urban Water**, v. 3, n. 4, p. 261–270, dez. 2001.

BEALE, R.; JACKSON, T. Pattern Recognition. In: **Neural Computing - An introduction**. New York: Taylor&Francis Group, 2010. p. 15–37.

BUCHBERGER, S. G.; NADIMPALLI, G. Leak Estimation in Water Distribution Systems by Statistical Analysis of Flow Readings. v. 130, n. 4, p. 321–329, 2004.

CASTRO, F. DE; CASTRO, M. C. DE. Capítulo 5 Radial Basis Function. In: **Redes Neurais Artificiais**. Porto Alegre: Pontífica Universidade Católica de Rio Grande do Sul, [s.d.].

COLOMBO, A. F.; LEE, P.; KARNEY, B. W. A selective literature review of transient-based leak detection methods. **Journal of Hydro-environment Research**, v. 2, n. 4, p. 212–227, abr. 2009.

COLOMBO, F. **Calibração de modelos hidráulicos de redes de abastecimento de água de sistemas reais admitindo vazamentos**. São Carlos: Universidade de São Paulo, 2007.

COVAS, D.; RAMOS, H. Case Studies of Leak Detection and Location in Water Pipe Systems by Inverse Transient Analysis. **Journal of Water Resources Planning and Management**, v. 136, n. 2, p. 248–257, 2010.

COVAS, D.; RAMOS, H.; ALMEIDA, A. B. DE. Standing Wave Difference Method for Leak Detection in Pipeline Systems. **Journal of Hydraulic Engineering**, v. 131, n. 12, p. 1106–1116, 2005.

DUDA, R.; HART, P.; STORK, D. **Pattern Classification**. 2. ed. New York: Willey-Interscience, 2001.

FACELI, K. et al. **Inteligência artificial. Uma abordagem de aprendizado de máquina**. Rio de Janeiro: LTC, 2011.

FERRANTE, M.; BRUNONE, B. Pipe system diagnosis and leak detection by unsteady-state tests. 1. Harmonic analysis. **Advances in Water Resources**, v. 26, n. 1, p. 95–105, jan. 2003.

FERRANTE, M.; BRUNONE, B. Pipe system diagnosis and leak detection by unsteady-state tests. 2. Wavelet analysis. **Advances in Water Resources**, v. 26, n. 1, p. 107–116, jan. 2003.

FERRANTE, M.; BRUNONE, B.; MENICONI, S. Wavelets for the Analysis of Transient Pressure Signals for Leak Detection. **Journal of hydraulic engineering**, v. 133, n. 11, p. 1274–1282, 2007.

FERRANTE, M.; BRUNONE, B.; MENICONI, S. Leak detection in branched pipe systems coupling wavelet analysis and a Lagrangian model. **Journal of Water Supply: Research and Technology—AQUA**, v. 58, n. 2, p. 95, mar. 2009.

GRAPS, A. An introduction to wavelets. **IEE Computational Science and Engineering**, v. 2, n. 2, 1995.

HOLTE, R. C. Very Simple Classification Rules Perform Well on Most Commonly Used Datasets. **Machine Learning**, v. 11, p. 63–90, 1 abr. 1993.

KAPELAN, Z.; SAVIC, D.; WALTERS, G. A hybrid inverse transient model for leakage detection and roughness calibration in pipe networks. **JOURNAL OF HYDRAULIC RESEARCH**, v. 41, n. 5, p. 481–492, 2003.

LIGGETT, J. A.; CHEN, L.-C. Inverse transient analysis in pipe networks. **Journal of hydraulic engineering**, v. 120, n. 8, p. 934–955, 1994.

MALLAT, S. **A Wavelet Tour of Signal Processing**. 2. ed. London: Academic Press, 1999.

MENICONI, S. et al. Potential of transient tests to diagnose real supply pipe systems : What can be done with a single extemporaneous test. **Journal of Water Resources Planning and Management**, v. 137, n. 2, p. 238–241, 2011.

MEYER, Y. Time-Frequency / Time-Scale Analysis. In: **Wavelet Analysis and its applications**. London: Academic Press, 1999. v. Volume 10.

MISIUNAS, D. **Burst Detection and Location in Pipelines and Pipe Networks with application in water distribution systems**. Lund: Lund University, 2003.

MISIUNAS, D. et al. Pipeline Break Detection Using Pressure Transient Monitoring. **Journal of Water Resources Planning and Management**, v. 131, n. 4, p. 316–325, 1 jul. 2005.

MOUNCE, S. R.; BOXALL, J. B.; MACHELL, J. Development and verification of an online artificial intelligence system for detection of bursts and other abnormal flows. **Journal of Water Resources Planning and Management**, v. 136, n. 3, p. 309–318, 2010.

MPESHA, W.; GASSMAN, S. L.; CHAUDHRY, M. H. Leak detection in pipes by frequency response method. **Journal of hydraulic engineering**, v. 127, n. 2, 2001.

NIXON, W. et al. Range of Validity of the Transient Damping Leakage. **Journal of hydraulic engineering**, v. 132, n. 9, p. 944–957, 2006.

PALAU, C. V.; ARREGUI, F. J.; CARLOS, M. Burst Detection in Water Networks Using Principal Component Analysis. **Journal of Water Resoucer planning and management**, v. 138, n. 1, p. 47–54, 2012.

PILCHER, R. **Leak location and repair guidance notes and the never ending war against leakage**Water Loss. **Anais...**Bucareste, Romênia: 2007

POULAKIS, Z.; VALOUGEORGIS, D.; PAPADIMITRIOU, C. Leakage detection in water pipe networks using a Bayesian probabilistic framework. **Probabilistic Engineering Mechanics**, v. 18, n. 4, p. 315–327, out. 2003.

PROAKIS, J. G.; MANOLAKIS, D. G. **Digital Signal Processing**. 4. ed. UpperSaddleRiver: Pearson Education, 2007.

PUDAR, R. S.; LIGGETT, J. A. Leaks in Pipe Networks. **Journal of Hydraulic Engineering**, v. 118, n. 7, p. 1031–1046, jul. 1992.

PUUST, R. et al. A review of methods for leakage management in pipe networks. **Urban Water Journal**, v. 7, n. 1, p. 25–45, fev. 2010.

QUINLAN, J. R. **C4.5: programs for machine learning**. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993.

REIS, L. F.; PORTO, R.; CHAUDHRY, F. Optimal location of control valves in pipe networks by genetic algorithm. **Journal of Water Resources Planning and Management**, v. 123, n. 6, p. 317–326, 1997.

REZENDE, S. O. **Sistemas inteligentes. Fundamentos e Aplicações**. Barueri: Manole, 2003.

ROMANO, M.; KAPELAN, Z.; SAVIC, D. **Bayesian-based online burst detection in water distribution systems**10th International Conference on computing and control for the Water Industry. **Anais...**Sheffield: 2009

ROMANO, M.; KAPELAN, Z.; SAVIC, D. **Anomaly Detection Based in Bayesian Inference**GBPCT/GB2010/000961, , 2010.

ROMANO, M.; KAPELAN, Z.; SAVIC, D. Automated Detection of pipe burst and other events in water distribution systems. **Journal of Water Resources Planning and Management**, 2012.

ROMANO, M.; KAPELAN, Z.; SAVIC, D. **Testing of the system for detection of pipe bursts and other events in a uk water distribution system**14th Water Distribution System Analysis Conference. **Anais...**Adelaide: 2012

SALDARRIAGA, J. G.; FUENTES, D. A. A.; GALVIS, L. F. C. **Implementation of the Hydraulic Transient and Steady Oscillatory Flow with Genetic Algorithms for Leakage Detection in Real Water Distribution Networks**Water Distribution Systems Analysis Symposium 2006Reston, VAASCE, , 13 mar. 2006.

SNSA, S. N. DE S. A. **Sistema Nacional de Informações sobre Saneamento: diagnóstico dos serviços de água e esgotos - 2008**. Brasília: MCIDADES, 2010.

SOARES, A. **Calibração de modelos de redes de distribuição de água para abastecimento considerando vazamentos e demandas dirigidas pela pressão**. São Carlos: Universidade de São Paulo, 2003.

SOARES, A. **Calibração e Detecção de Vazamentos em Modelos de Sistemas Hidráulicos no Escoamento Transitório**. São Carlos: Universidade de São Paulo, 2007.

SOARES, A. K.; COVAS, D. I. C.; REIS, L. F. R. Leak detection by inverse transient analysis in an experimental PVC pipe system. **Journal of Hydroinformatics**, v. 13, n. 2, p. 153, abr. 2011.

THEODORIDIS, S.; KOUTROUMBAS, K. **Pattern recognition**. San Diego: Academic Press, 1999.

VITKOVSKY, J. P.; SIMPSON, A. R.; LAMBERT, M. F. Leak detection and calibration using transient and genetic algorithms. **Journal of Water Resources Planning and Management**, v. 126, n. 4, p. 262–265, 2000.

WANG, X. et al. Leak Detection in Pipelines using Damping of Fluid Transients. **Journal of hydraulic engineering**, v. 128, n. 7, p. 697–711, 2002.

WLTS, W. District Metered Areas Guidance Notes. 2007.

WU, Z. Y. Unified parameter optimisation approach for leakage detection and extended-period simulation model calibration. **Urban Water Journal**, v. 6, n. 1, p. 53–67, mar. 2009.

WU, Z. Y.; SAGE, P.; TURTLE, D. Pressure-Dependent Leak Detection Model and Its Application to a District Water System. **Journal of Water Resources Planning and Management**, v. 136, n. 1, p. 116–128, 2010.

YE, G.; PH, D.; FENNER, R. A. Kalman Filtering of Hydraulic Measurements for Burst Detection in Water Distribution Systems. **Journal of pipeline systems engineering and practice**, v. 2, n. 1, p. 14–22, 2011.

APÊNDICE A. HISTOGRAMAS DOS VALORES DAS CARACTERÍSTICAS EXTRAÍDAS DOS SINAIS PARA TODOS OS SENSORES

Frequência (quantidade de exemplos). Sensor P03

CLASSE
 COM
 SEM

Frequência (quantidade de exemplos). Sensor P04

CLASSE
 COM
 SEM

Frequência (quantidade de exemplos). Sensor P05

CLASSE
 COM
 SEM

Frequência (quantidade de exemplos). Sensor P06

CLASSE
 COM
 SEM

Frequência (quantidade de exemplos). Sensor P07

Frequência (quantidade de exemplos). Sensor P08

Frequência (quantidade de exemplos). Sensor P09

Frequência (quantidade de exemplos). Sensor P11

Frequência (quantidade de exemplos). Sensor P12

CLASSE
☐ COM
☐ SEM

Frequência (quantidade de exemplos). Sensor P13

CLASSE
☐ COM
☐ SEM

Frequência (quantidade de exemplos). Sensor P14

Frequência (quantidade de exemplos). Sensor P15

Frequência (quantidade de exemplos). Sensor P16

Frequência (quantidade de exemplos). Sensor Q01

Frequência (quantidade de exemplos). Sensor Q02

Frequência (quantidade de exemplos). Sensor Q03

**APÊNDICE B. DESEMPENHO DO SISTEMA DE RECONHECIMENTO DE
PADRÕES PROPOSTO, RESULTADOS INDIVIDUAIS DA VALIDAÇÃO
CRUZADA**

Sensor	Iteraç.	VC	FS	FC	VS	% acerto	AAC
P01	1	29	2	0	31	96.77%	0.965
	2	28	3	2	29	91.94%	0.969
	3	29	2	1	30	95.16%	0.960
	4	30	1	2	29	95.16%	0.977
	5	30	1	0	31	98.39%	1.000
	6	26	5	1	30	90.32%	0.967
	7	27	4	6	25	83.87%	0.920
	8	30	1	0	31	98.39%	0.999
	9	28	3	0	31	95.16%	0.975
	10	28	3	3	28	90.32%	0.939
P02	1	31	0	0	31	100.00%	1.000
	2	31	0	0	31	100.00%	1.000
	3	31	0	1	30	98.39%	1.000
	4	30	1	2	29	95.16%	0.981
	5	30	1	2	29	95.16%	0.996
	6	31	0	1	30	98.39%	0.984
	7	31	0	1	30	98.39%	0.982
	8	31	0	1	30	98.39%	0.999
	9	30	1	1	30	96.77%	0.983
	10	29	2	2	29	93.55%	0.979
P03	1	31	0	0	31	100.00%	1.000
	2	31	0	0	31	100.00%	1.000
	3	31	0	2	29	96.77%	0.968
	4	30	1	2	29	95.16%	0.979
	5	30	1	1	30	96.77%	0.981
	6	30	1	1	30	96.77%	0.982
	7	29	2	2	29	93.55%	0.966
	8	31	0	0	31	100.00%	1.000
	9	29	2	1	30	95.16%	0.997
	10	28	3	1	30	93.55%	0.980
P04	1	31	0	0	31	100.00%	1.000
	2	31	0	0	31	100.00%	1.000
	3	31	0	2	29	96.77%	0.983
	4	30	1	1	30	96.77%	0.997
	5	29	2	1	30	95.16%	0.980
	6	31	0	1	30	98.39%	0.982
	7	30	1	1	30	96.77%	0.966
	8	30	1	0	31	98.39%	0.984
	9	29	2	0	31	96.77%	0.999
	10	29	2	1	30	95.16%	0.980

Sensor	Iteraç.	VC	FS	FC	VS	% acerto	AAC
P05	1	29	2	2	29	93.55%	0.980
	2	30	1	0	31	98.39%	0.983
	3	30	1	3	28	93.55%	0.960
	4	31	0	1	30	98.39%	0.998
	5	30	1	2	29	95.16%	0.966
	6	29	2	3	28	91.94%	0.947
	7	29	2	4	27	90.32%	0.939
	8	29	2	0	31	96.77%	0.997
	9	30	1	2	29	95.16%	0.993
	10	30	1	2	29	95.16%	0.979
P06	1	30	1	0	31	98.39%	0.998
	2	31	0	0	31	100.00%	1.000
	3	31	0	2	29	96.77%	0.983
	4	29	2	1	30	95.16%	0.981
	5	30	1	2	29	95.16%	0.980
	6	31	0	1	30	98.39%	0.983
	7	30	1	1	30	96.77%	0.982
	8	29	2	1	30	95.16%	0.981
	9	29	2	1	30	95.16%	0.981
	10	27	4	2	29	90.32%	0.944
P07	1	30	1	0	31	98.39%	0.998
	2	31	0	0	31	100.00%	1.000
	3	30	1	2	29	95.16%	0.982
	4	28	3	0	31	95.16%	0.963
	5	30	1	1	30	96.77%	0.981
	6	30	1	1	30	96.77%	0.967
	7	30	1	1	30	96.77%	0.981
	8	30	1	0	31	98.39%	0.983
	9	31	0	2	29	96.77%	0.997
	10	29	2	1	30	95.16%	0.978
P08	1	31	0	1	30	98.39%	0.983
	2	31	0	0	31	100.00%	1.000
	3	31	0	2	29	96.77%	0.967
	4	28	3	2	29	91.94%	0.974
	5	29	2	0	31	96.77%	0.983
	6	27	4	2	29	90.32%	0.931
	7	30	1	2	29	95.16%	0.979
	8	29	2	2	29	93.55%	0.961
	9	29	2	2	29	93.55%	0.978
	10	30	1	1	30	96.77%	0.981
P09	1	30	1	1	30	96.77%	0.998
	2	31	0	0	31	100.00%	1.000
	3	31	0	2	29	96.77%	0.967
	4	28	3	1	30	93.55%	0.994
	5	29	2	2	29	93.55%	0.964
	6	29	2	1	30	95.16%	0.950
	7	31	0	3	28	95.16%	0.982
	8	29	2	0	31	96.77%	0.998
	9	30	1	1	30	96.77%	0.982
	10	28	3	2	29	91.94%	0.976

Sensor	Iteraç.	VC	FS	FC	VS	% acerto	AAC
P11	1	30	1	0	31	98.39%	0.998
	2	29	2	0	31	96.77%	0.996
	3	31	0	2	29	96.77%	0.967
	4	30	1	0	31	98.39%	0.982
	5	30	1	1	30	96.77%	0.982
	6	30	1	1	30	96.77%	0.966
	7	30	1	1	30	96.77%	0.996
	8	31	0	1	30	98.39%	0.983
	9	29	2	0	31	96.77%	0.996
	10	29	2	2	29	93.55%	0.977
P12	1	31	0	0	31	100.00%	1.000
	2	31	0	0	31	100.00%	1.000
	3	31	0	2	29	96.77%	0.984
	4	29	2	2	29	93.55%	0.980
	5	30	1	1	30	96.77%	0.980
	6	30	1	1	30	96.77%	0.967
	7	30	1	1	30	96.77%	0.982
	8	30	1	0	31	98.39%	0.999
	9	30	1	0	31	98.39%	0.999
	10	29	2	1	30	95.16%	0.981
P13	1	30	1	0	31	98.39%	0.982
	2	30	1	0	31	98.39%	0.983
	3	31	0	0	31	100.00%	1.000
	4	30	1	0	31	98.39%	0.983
	5	29	2	2	29	93.55%	0.980
	6	27	4	1	30	91.94%	0.946
	7	29	2	2	29	93.55%	0.948
	8	31	0	0	31	100.00%	1.000
	9	31	0	0	31	100.00%	1.000
	10	31	0	1	30	98.39%	1.000
P14	1	31	0	0	31	100.00%	1.000
	2	31	0	0	31	100.00%	1.000
	3	28	3	1	30	93.55%	0.951
	4	28	3	1	30	93.55%	0.978
	5	31	0	2	29	96.77%	0.998
	6	27	4	2	29	90.32%	0.930
	7	29	2	4	27	90.32%	0.971
	8	30	1	1	30	96.77%	0.998
	9	30	1	2	29	95.16%	0.979
	10	31	0	2	29	96.77%	0.981
P15	1	30	1	1	30	96.77%	0.997
	2	30	1	1	30	96.77%	0.996
	3	31	0	2	29	96.77%	0.981
	4	30	1	1	30	96.77%	0.997
	5	30	1	0	31	98.39%	0.999
	6	27	4	1	30	91.94%	0.918
	7	30	1	3	28	93.55%	0.959
	8	29	2	0	31	96.77%	0.998
	9	30	1	1	30	96.77%	0.982
	10	30	1	1	30	96.77%	0.981

Sensor	Iteraç.	VC	FS	FC	VS	% acerto	AAC
P16	1	30	1	1	30	96.77%	0.982
	2	31	0	0	31	100.00%	1.000
	3	30	1	3	28	93.55%	0.945
	4	31	0	2	29	96.77%	0.980
	5	31	0	1	30	98.39%	0.999
	6	29	2	2	29	93.55%	0.932
	7	29	2	4	27	90.32%	0.943
	8	31	0	0	31	100.00%	1.000
	9	29	2	0	31	96.77%	0.968
	10	29	2	2	29	93.55%	0.964
Q01	1	28	3	1	30	93.55%	0.974
	2	27	4	2	29	90.32%	0.938
	3	27	4	2	29	90.32%	0.968
	4	30	1	4	27	91.94%	0.972
	5	29	2	3	28	91.94%	0.944
	6	22	9	2	29	82.26%	0.932
	7	30	1	4	27	91.94%	0.968
	8	29	2	1	30	95.16%	0.977
	9	31	0	3	28	95.16%	0.978
	10	29	2	5	26	88.71%	0.977
Q02	1	28	3	4	27	88.71%	0.950
	2	29	2	4	27	90.32%	0.959
	3	26	5	0	31	91.94%	0.944
	4	30	1	3	28	93.55%	0.980
	5	27	4	1	30	91.94%	0.961
	6	27	4	2	29	90.32%	0.950
	7	27	4	3	28	88.71%	0.937
	8	29	2	2	29	93.55%	0.963
	9	30	1	2	29	95.16%	0.993
	10	28	3	3	28	90.32%	0.975
Q03	1	29	2	6	25	87.10%	0.956
	2	25	6	2	29	87.10%	0.928
	3	24	7	8	23	75.81%	0.802
	4	26	5	6	25	82.26%	0.888
	5	26	5	5	26	83.87%	0.915
	6	21	10	2	29	80.65%	0.900
	7	27	4	5	26	85.48%	0.892
	8	29	2	3	28	91.94%	0.933
	9	29	2	6	25	87.10%	0.906
	10	28	3	4	27	88.71%	0.935

APÊDICE C. CURVA ROC DO CLASSIFICADOR NO SISTEMA DE RECONHECIMENTO DE PADRÕES PROPOSTO (TODOS OS SENSORES)

P01

P02

P03

P04

P05

P06

P07

P08

P09

P11

P12

P13

P14

P15

P16

Q01

Q02

Q03