UNIVERSIDADE FEDERAL FLUMINENSE

FILIPE VENTURA MUGGIATI

MINERAÇÃO DE DADOS OPERATIVOS DE USINAS HIDRELÉTRICAS

NITERÓI

2019

UNIVERSIDADE FEDERAL FLUMINENSE

FILIPE VENTURA MUGGIATI

MINERAÇÃO DE DADOS OPERATIVOS DE USINAS HIDRELÉTRICAS

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Computação da Universidade Federal Fluminense como requisito parcial para a obtenção do Grau de Mestre em Computação. Área de concentração: Computação Científica e Sistemas de Potência

Orientadores:
Milton Brown Do Coutto Filho

Julio Cesar Stacchini de Souza

NITERÓI

2019

Ficha catalográfica automática - SDC/BEE
Gerada com informações fornecidas pelo autor

M951m Muggiati, Filipe Ventura
Mineração de Dados Operativos de Usinas Hidrelétricas / Filipe Ventura Muggiati ; Milton Brown Do Coutto Filho, orientador ; Julio Cesar Stacchini de Souza, coorientador. Niterói, 2019.
81 f. : il.

Dissertação (mestrado)-Universidade Federal Fluminense, Niterói, 2019.

DOI: http://dx.doi.org/10.22409/PGC.2019.m.04335039980

1. Mineração de dados (Computação). 2. Usina hidrelétrica. 3. Produção intelectual. I. Do Coutto Filho, Milton Brown, orientador. II. Souza, Julio Cesar Stacchini de, coorientador. III. Universidade Federal Fluminense. Instituto de Computação. IV. Título.

CDD -

Bibliotecária responsável: Fabiana Menezes Santos da Silva - CRB7/5274

FILIPE VENTURA MUGGIATI

MINERAÇÃO DE DADOS OPERATIVOS DE USINAS HIDRELÉTRICAS

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Computação da Universidade Federal Fluminense como requisito parcial para a obtenção do Grau de Mestre em Computação. Área de concentração: Computação Científica e Sistemas de Potência

Aprovada em março de 2019.

BANCA EXAMINADORA

Prof. Milton Brown Do Coutto Filho, D.Sc. - Orientador
Universidade Federal Fluminense

Prof. Julio Cesar Stacchini de Souza, D.Sc. - Orientador
Universidade Federal Fluminense

Prof. Alexandre Plastino de Carvalho, D.Sc.
Universidade Federal Fluminense

Profª. Marley Maria B. R. Vellasco, Ph.D.
Pontifícia Universidade Católica do Rio de Janeiro

Niterói
2019

Aos meus queridos pais, Maria Cristina e Ruy.

Agradecimentos

À Itaipu Binacional, pelo fundamental apoio para a realização deste trabalho.

Ao Engenheiro Celso Villar Torino, pelo incentivo e confiança depositada neste projeto.

Aos orientadores Milton Brown Do Coutto Filho e Julio Cesar Stacchini de Souza, pelos ensinamentos, sugestões e revisões feitas durante toda a elaboração deste trabalho.

Aos colegas Paulo Zanelli, Jefferson Stumpf e Bernardo Morcelli, pelo apoio na realização dos testes e casos de uso da metodologia proposta.

Ao colega Felipe Trevisan e ao estagiário Gustavo Scalco pelas sugestões e testes da metodologia e da ferramenta computacional.

Aos demais colegas da minha equipe, Paulo Benites, Paulo Neis, Rafael Deitos, Teresinha Arnauts, Carolina Villasanti, Edgar Fernandez e Jose Flores, pelas sugestões feitas e pelo suporte durante minha ausência.

A todos os professores da UFF que de alguma forma contribuíram para a realização deste trabalho e meu curso de mestrado, em especial ao professor Alexandre Plastino pelos ensinamentos na disciplina de mineração de dados.

Aos colegas que administram a área em que trabalho na Itaipu, Cristiane Pimenta, Christian Le Bourlegat e Elias Alberto da Silva, e à equipe da secretaria da pós-gradução, Teresa Cancela e Helio Augusto, pelo apoio na organização de todo o processo necessário para minha realização do curso.

Aos amigos Maria e Samuel, pela hospitalidade durante minha estadia no Rio de Janeiro.

À minha noiva Adrieli, aos meus pais, Maria Cristina e Ruy, e aos meus irmãos, Bruna e Juliano, pelo apoio e incentivo durante todo o curso.

Resumo

Esta Dissertação apresenta uma metodologia baseada na aplicação de técnicas de mineração de dados para a detecção de anomalias em séries temporais de grandezas físicas que tenham sido registradas por um sistema de supervisão e controle de uma usina hidrelétrica. A metodologia proposta pode ser utilizada por especialistas que busquem a solução de um conjunto amplo de problemas relacionados à supervisão de processos industriais. Inicialmente, os dados registrados são filtrados, transformados e agregados em um formato compatível com algoritmos de mineração de dados. Em seguida, aplica-se um algoritmo de classificação não-supervisionado para realizar uma pré-rotulagem, reduzindo significativamente o trabalho do especialista na avaliação dos dados. Posteriormente, novos dados são classificados em tempo real usando um algoritmo supervisionado. Um software para identificação de anomalias em usinas hidroelétricas foi construído durante o desenvolvimento deste trabalho e testado empregando-se dados simulados, bem como outros correspondentes a situações reais. Os resultados obtidos são apresentados e discutidos ao longo da Dissertação.

Palavras-chave: Mineração de Dados, Séries Temporais, Detecção de Anomalias, Geração de Energia Elétrica, Supervisão de Processos Industriais.

Abstract

This work proposes a methodology based on data mining techniques, aiming at detecting anomalies in real time measurements stored as time series data, recorded by a supervisory control system of a hydroelectric power plant. The proposed methodology is developed to be used by specialists interested in solving problems related to the supervision of industrial processes. The recorded data are filtered, transformed and aggregated into a format compatible with data mining algorithms. Then, an unsupervised classifier is applied on the unlabeled data to perform a preliminary data labeling, significantly reducing the specialist task of data analysis. Thus, new data are classified in real time using a supervised algorithm. A software for the identification of anomalies in power plants was developed and tested using simulated and real data. The obtained results are presented and discussed.

Keywords: data-mining, time series, anomaly detection, electrical power generation, industrial process supervision.

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

API	:	Application Programming Interface;
DW	:	Data Warehouse;
ETL	:	Extration, Transformation and Load;
IHM	:	Interface Homem-Máquina;
KDD	:	Knowledge Discovery from Data;
KNN	:	k-nearest neighbors;
MLP	:	Multilayer Perceptron;
OLAP	:	Online Analytical Processing;
RNN	:	Replicator Neural Network;
SCADA	:	Supervisory Control and Data Acquisition;
SGBD	:	Sistema de Gerenciamento de Banco de Dados;
TC	:	Transformador de Corrente;
UTM	:	Unidade Terminal Mestre;
UTR	:	Unidade Terminal Remota;

Sumário

Capítulo 1

Introdução

1.1 Considerações Iniciais

A atividade de controle e supervisão realizada em processos industriais, em especial naqueles destinados à produção de energia elétrica, ocorre em cenários altamente especializados em que decisões complexas e de grande impacto precisam ser tomadas, muitas vezes, em curto intervalo de tempo. Assim, a capacidade de identificação automática de eventos relevantes, até mesmo os ainda não observados ou raros, pode ser muito útil para a detecção de anomalias e diagnóstico de falhas.

Anomalias referentes ao funcionamento de equipamentos de usinas geradoras de sistemas elétricos de potência são importantes eventos que ensejam perturbações, desde danos diversos aos equipamentos em si, até a interrupção no fornecimento de energia elétrica aos consumidores. Muitas destas anomalias podem ser detectadas através de regras baseadas nos princípios e limites físicos de funcionamento dos equipamentos. Para isso são utilizados, por exemplo, dispositivos de proteção que monitoram algumas grandezas, emitem alarmes, ou até mesmo interrompem seu funcionamento. Deve-se ter em conta que alguns processos são bastante complexos e que estes dispositivos de proteção tradicionais não resultam em ações satisfatórias em certas condições, pois muitas vezes a anomalia se caracteriza pela relação entre diversas grandezas do cenário em análise, e não apenas por seus valores absolutos. Além disso, a grande quantidade de equipamentos, existentes nos processos envolvidos na geração de energia, também dificulta a formulação de modelos para a identificação de anomalias.

Com a digitalização dos sistemas elétricos de potência, novas oportunidades para o desenvolvimento de ferramentas de suporte às equipes de operação têm surgido. A

capacidade de registrar dados digitais, coletados por diversos equipamentos, permitiu a criação de grandes bancos de dados históricos, dos quais uma grande quantidade de informações pode ser extraída a cada instante. Estes dados representam séries temporais de grandezas físicas, usualmente correspondentes a condições normais de operação.

Neste contexto, técnicas de mineração de dados e aprendizado de máquina podem ser utilizadas, pois permitem a análise e processamento (treinamento) de um grande conjunto de dados históricos, potencialmente capazes de identificar e correlacionar anomalias em um espaço de busca multidimensional.

1.2 Objetivos

Neste trabalho busca-se avaliar a utilização de técnicas de mineração de dados e aprendizado de máquina para expandir a capacidade humana de supervisão em usinas de produção de energia elétrica e definir uma metodologia para a aplicação destas técnicas de forma consistente. Para isso, foi realizado um estudo de caso na Usina de Itaipu com a utilização de diferentes tipos de modelos computacionais de mineração de dados, treinados com um conjunto histórico de dados, para avaliar a capacidade destes algoritmos em identificar e correlacionar anomalias da operação de sistemas elétricos de potência, visando fornecer às equipes de operação informações úteis para a tomada de decisão.

A abordagem proposta nesta Dissertação pode ser útil para a supervisão de outros tipos de processos industriais, tais como aqueles que registrem seus dados históricos de operação na forma de séries temporais.

Resumidamente, os objetivos específicos desta Dissertação são:

- avaliação dos resultados alcançados com a aplicação de algoritmos de mineração de dados para detecção de anomalias, utilizando métodos de treinamento supervisionados ou não, considerando que a rotulagem (supervisão) dos dados pode ser um limitador para a ampla utilização destas técnicas no cenário apresentado;
- elaboração de uma metodologia (tão ampla quanto possível) para uso de especialistas (de áreas diversas de conhecimento) interessados na solução de problemas relacionados à supervisão de processos industriais por meio de técnicas de mineração de dados;
- construção de uma ferramenta de suporte à operação, baseada na metodologia proposta, capaz de potencializar a experiência e a capacidade de análise dos operadores

de usinas geradoras de energia elétrica e, com isto, validar os resultados da aplicação desta metodologia em um ambiente industrial.

1.3 Metodologia de Pesquisa

Este trabalho foi desenvolvido utilizando uma metodologia de pesquisa baseada na teoria de *Design Science* [24]. Esta propõe uma sequência iterativa entre a especificação (feita de acordo com necessidades identificadas através de pesquisa realizada com os interessados), o desenvolvimento de um artefato (*software*) e a análise da aplicação deste *software* no seu ambiente de uso. Avalia-se como os usuários reagem à utilização do software para aprimorá-lo sucessivamente. A partir destas análises, pode ser feito o levantamento de hipóteses e teorias para produzir conhecimento científico e, por fim, um resultado final, capaz de atender às expectativas e objetivos dos *stakeholders* (partes interessadas na pesquisa) e patrocinadores.

Desta forma, importa afirmar que técnicos e engenheiros da operação de usinas geradoras de energia elétrica são os interessados (*stakeholders*) deste trabalho de pesquisa. Assim, inicialmente foram feitas entrevistas com estes profissionais para identificar potenciais problemas operativos, ilustrados neste trabalho. Após a realização de uma revisão da literatura sobre o tema e uma proposta de metodologia, partiu-se para a sua implementação computacional. O *software* desenvolvido foi aplicado a estudos de casos, referentes a problemas relatados por operadores da planta, e aprimorado de modo a atender às expectativas e necessidades dos *stakeholders*. Por fim, soluções encontradas foram generalizadas e relatadas na presente Dissertação.

1.4 Revisão Bibliográfica

O termo anomalia refere-se a aquilo que se desvia do comportamento esperado, condição fora do comum, sendo anômalo, irregular, apresentando desequilíbrio, falha, erro. Em 1980, Hawkins definiu uma anomalia como "uma observação tão diferente das outras observações que levanta suspeitas de que foi gerada por um mecanismo diferente"[13]. Desde então, a detecção de anomalias foi tema de diversos artigos, pesquisas e livros já publicados em diversas áreas de pesquisa, como a estatística, teoria da informação, aprendizado de máquina e mineração de dados.

Chandola et al. (2009) [3] publicou uma pesquisa abrangente sobre as diversas classi-

ficações de anomalias e as diversas técnicas propostas para detectá-las. Em sua pesquisa, Chandola apresenta diversas aplicações para a detecção de anomalias, como a detecção de fraudes, ataques em redes de computadores, saúde e proteção de danos na indústria. Apresenta também as definições de contextualização e coletividade de anomalias e como estas características influenciam no tratamento do problema e, ainda, diversos métodos utilizados para a detecção de anomalias, como métodos baseados em modelos estatísticos, agrupamentos (*clustering*) e classificadores.

Widmer et al. (1996) [23] apresentou métodos para o tratamento de mudanças do contexto ao longo do tempo para o treinamento de algoritmos de mineração de dados e aprendizado de máquinas para atividades como a detecção de anomalias em fluxo continuo de dados.

Hawkins et al. (2002) [14] propôs a utilização de redes neurais replicantes (*replicator neural network*), até então utilizadas para compressão de imagens, para a detecção de anomalias de forma não supervisionada, considerando a premissa de que os dados normais representam a grande maioria do conjunto de treinamento.

Liu et al. (2008) [18] publicou trabalho demonstrando a capacidade de conjuntos de árvores de decisão (florestas), construídas aleatoriamente, identificarem anomalias, também de forma não supervisionada, através da avaliação do comprimento do caminho percorrido na árvore na avaliação de cada instância.

Dau et al. (2014) [7] propôs, baseado no trabalho de Hawkins et. al. [14] a utilização de redes neurais replicantes de apenas uma classe para a detecção de anomalias, o que permitiria definir automaticamente um limiar para classificar anomalias. Em seu trabalho, Dau também demonstrou que a utilização de apenas uma camada oculta levaria a bons resultados e simplificaria muito o processo, especialmente na definição dos parâmetros da rede.

Goldstein et al. (2014) [9] verificou a existência de uma lacuna entre os dados resultantes de processos reais, como sensores de equipamentos industriais, e as estruturas dos algoritmos existentes para detecção de anomalias. Goldstein apontou a necessidade de realizar transformações nestes dados, na forma de agregações, para a criação das chamadas visualizações de dados (*data views*) com a estrutura adequada para a aplicação dos algoritmos. Para o caso de séries temporais, Goldstein aponta que estas agregações poderiam ser feitas em função do tempo.

Bezerra et al. (2012) [1] publicou trabalho avaliando a utilização de técnicas de mi-

neração de dados em um histórico de registros de um sistema de controle e supervisão de uma usina hidrelétrica. Em seu trabalho, Bezerra analisa o uso de métodos estatísticos para revelar aspectos relevantes de grandezas de estado, como alarmes, comandos e set-points, e grandezas analógicas, como a frequência elétrica. Além disso, foram também analisadas regras de associação, no intuito de identificar relacionamento entre o acionamento de determinados alarmes do sistema. Por fim, foi feita a construção de uma árvore de decisão para descrever o comportamento da geração em função da descarga de água nas turbinas.

Bezerra conclui em seu trabalho que as técnicas de mineração de dados apresentaram resultados satisfatórios para os problemas ali propostos e poderiam ser usadas em análises pós-operativas e em tempo real. Porém, Bezerra destaca as dificuldades que as características multi-disciplinares, necessárias para a aplicação destas técnicas, representam. Ou seja, o analista deve conhecer tanto do negócio quando das técnicas. Por fim, Bezerra também pontua a ausência de dados rotulados como mais uma dificuldade.

Apesar de compartilhar de técnicas de análise e fontes de dados semelhantes com a pesquisa de Bezerra, o trabalho aqui proposto não se dedica a solução de problemas específicos, mas a elaboração de uma metodologia de uso amplo e para uso por especialistas do negócio. Neste sentido, justamente as dificuldades apontadas por Bezerra (características multi-disciplinares e ausência de dados rotulados) representam grandes desafios para este trabalho.

1.5 Estrutura da Dissertação

Esta Dissertação está estruturada da seguinte forma: o segundo capítulo apresenta uma contextualização sobre a atividade de supervisão de uma planta de geração de energia elétrica, suas principais características, dificuldades, desafios. O terceiro capítulo diz respeito à fundamentação teórica e apresentação dos conceitos aqui utilizados, tais como: descoberta de Conhecimento, mineração de dados e algoritmos para mineração de dados. O quarto capítulo descreve as principais características do problema pesquisado e uma proposta de metodologia para a utilização de técnicas de mineração para a análise de séries temporais históricas de dados operativos de uma usina hidrelétrica. O quinto capítulo apresenta alguns estudos de casos e os resultados obtidos com a aplicação da metodologia proposta na usina hidrelétrica de Itaipu. Por fim, o sexto capítulo reúne as conclusões gerais do trabalho e propõe alguns temas para pesquisas futuras.

Capítulo 2

Supervisão de Plantas de Geração Hidrelétrica

2.1 Introdução

Inegavelmente, a energia elétrica se tornou uma das principais formas de energia utilizadas pela sociedade moderna. Exigentes padrões de qualidade e continuidade de fornecimento tornam complexas as tarefas de implantação, manutenção e monitoramento dos atuais sistemas elétricos de geração.

Figura 2.1: Usina Hidrelétrica de Itaipu.

Usinas hidrelétricas são instalações destinadas à produção de energia elétrica através da conversão da energia potencial gravitacional da água em energia cinética e, posterior-

mente, em energia elétrica. A Figura 2.1 apresenta uma imagem da usina de Itaipu, que será apresentada em detalhes na Seção 5.3. Neste tipo de usina, a água que atravessa os chamados condutos forçados movimenta uma turbina que, por sua vez, está conectada mecanicamente (através de um eixo) a um gerador e movimenta um elemento girante (rotor), mostrados na Figura 2.2 e localizados na casa de força, como ilustra a Figura 2.3. A movimentação deste elemento cria um campo magnético girante capaz de induzir uma corrente elétrica que é levada aos centros de consumo por meio de linhas de transmissão.

Figura 2.2: Turbina à esquerda, eixo ao centro e rotor à direita.

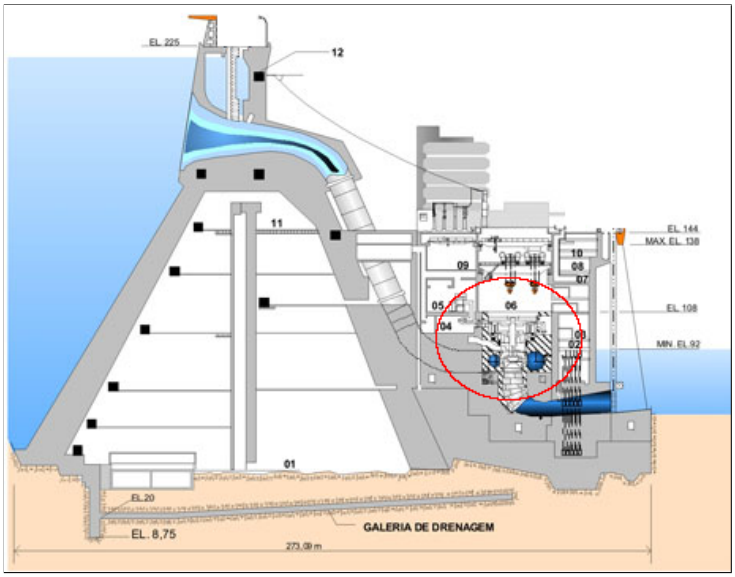

Figura 2.3: Localização da casa de força na usina de Itaipu.

Evidentemente, trata-se de um processo complexo, composto por diversos subsistemas que operam em conjunto para a produção de energia dentro dos padrões de qualidade e confiabilidade desejados. Cada etapa descrita anteriormente é composta por um grande conjunto de equipamentos e dispositivos que trabalham ininterruptamente durante longos períodos. Desta forma, a supervisão do processo de produção de energia elétrica torna-se naturalmente uma importante tarefa a ser realizada nas usinas.

Entre as diversas grandezas associadas à energia elétrica produzida, destacam-se a frequência elétrica e a tensão, proporcionais à frequência mecânica do movimento da turbina pela passagem da água e à corrente de campo utilizada para induzir o campo magnético. Além da frequência e da tensão, também são de interesse as potências ativa e reativa (integradas no tempo, resultam em energia). Estas grandezas são especialmente importantes, pois muitos dos processos e subsistemas que serão abordados mais adiante neste trabalho as supervisionam e nelas atuam, direta ou indiretamente.

Neste capítulo, apresenta-se brevemente uma descrição da atividade de supervisão de uma planta de geração de energia elétrica, com suas principais características, dificuldades e desafios.

2.2 Supervisão e Controle da Geração

O termo supervisão usualmente denota uma ampliação na capacidade de observação humana, sendo muito utilizado no contexto industrial para indicar que o comportamento de um sistema (como um todo ou em parte) está sendo acompanhado. Considerando a complexidade de alguns sistemas, como por exemplo o que representa a operação de uma planta de produção de energia elétrica, ferramentas computacionais são necessárias tanto no processo de aquisição dos dados (geograficamente distribuídos ou fisicamente inacessíveis), como na obtenção de informações úteis à tomada de decisão.

Cabe ressaltar que a atividade de supervisão de um sistema elétrico de potência muitas vezes está inserida em ambientes de características conservadoras, de certa forma resistentes à adoção de novas tecnologias que ainda não tenham sido completamente absorvidas em tais ambientes. Isto decorre da grande responsabilidade que recai sobre os operadores de sistemas de potência, em termos das decisões e ações a serem por eles tomadas, especialmente em cenários de operação com anormalidades. Desta forma, se coloca como um grande desafio a adoção de tecnologias inovadoras de forma segura e potencialmente capazes de melhorar a capacidade de supervisão da operação.

Além disso, tecnologias que possam auxiliar a equipe de supervisão a ter um conhecimento mais aprofundado das condições de operação dos equipamentos podem não só evitar eventuais falhas, decorrentes de anomalias, como também auxiliar na tomada de decisão das ações direcionadas à continuidade e qualidade do processo de produção de energia como um todo.

As atividades de supervisão e controle em usinas hidrelétricas eram, inicialmente,

realizadas através de um sistema analógico, formado por equipamentos e dispositivos eletromecânicos. Em muitas usinas, este sistema ainda se encontra operacional e pode, como é o caso da Itaipu Binacional, ser usado como um sistema de retaguarda para o atual sistema digital que, após sucessivos trabalhos de modernização, foi definido como o sistema de controle e supervisão primário desta usina.

Para auxiliar na tarefa de supervisão de usinas hidrelétricas, um sistema computacional, chamado SCADA (*Supervisory Control and Data Acquisition*), pode ser utilizado. Este sistema possui características muito particulares que o diferenciam de sistemas de monitoramento e controle tradicionais, por apresentarem uma capacidade de controle limitada e destinada a processos distribuídos [2]. O termo *supervisório* exerce um papel significativo na definição do sistema, pois denota justamente a atividade de monitorar o processo e apenas definir objetivos (*setpoints*) de controle, que são executados por dispositivos locais. A Figura 2.4 mostra a Sala de Controle Centralizado da Usina de Itaipu, onde é possível observar o sistema de controle analógico, em painéis brancos ao fundo, e o sistema de controle digital, operando a partir dos computadores das mesas no centro da sala.

Figura 2.4: Sala de Controle Central da usina de Itaipu.

Estes sistemas de supervisão e controle possuem, tradicionalmente, uma arquitetura em camadas, composto por uma unidade terminal mestre (UTM) e diversas unidades terminais remotas (UTRs). As UTRs contêm o *hardware* necessário para realizar as medidas físicas e atuar no processo. Já a UTM recebe todos os valores medidos pelas UTRs, calcula os sinais de controle e envia para os atuadores nas unidades remotas. A UTM normalmente é composta por *hardware* tradicional que, apesar de ter algumas particularidades para operar em tempo real, compartilha de muitas características dos sistemas computacionais tradicionais [2].

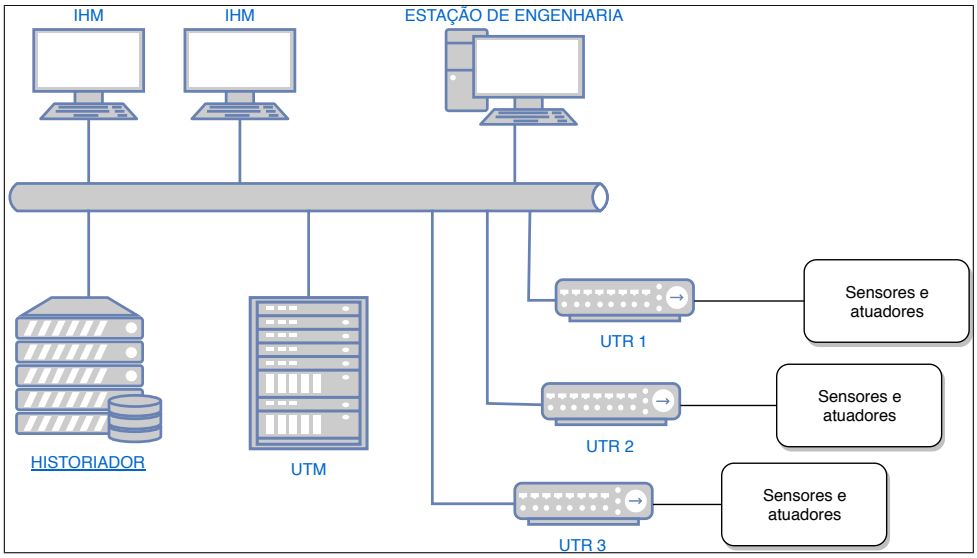

Figura 2.5: Arquitetura típica de um sistema SCADA.

Na Figura 2.5 é possível observar a arquitetura típica de um sistema SCADA, contendo a UTM, os dispositivos de campo, interfaces homem-máquina e o servidor de registro de dados históricos, que será detalhado a seguir.

Como a UTM é composta de um sistema computacional tradicional, ela pode ser integrada a um sistema de armazenamento de dados históricos, chamado de *historiador*, que funciona como um grande banco de dados de séries temporais. Isto permite que todas as medidas físicas e comandos, enviados aos atuadores ao longo do tempo para controle do processo, sejam registrados.

Historiadores comerciais apresentam ferramentas específicas para a visualização dos seus dados armazenados. Além disso, possuem algumas interfaces computacionais (*API*) para acesso a estes dados por outros aplicativos, como driver JDBC para aplicações Java, bibliotecas para as linguagens C/C++ e serviço web (REST).

Outra característica importante dos historiadores é a função de compressão de dados, aplicada para reduzir o espaço em disco necessário para o armazenamento de informações

redundantes. Seu funcionamento é feito através da definição de intervalos de valores relativos ao último valor registrado (banda morta), em que uma nova amostra só será registrada quando for recebido um valor fora deste intervalo. Assim, quando uma grandeza apresenta um comportamento suave, sem variações bruscas, apenas algumas amostras são registradas. Porém, ao apresentar um comportamento mais brusco, com grandes variações, muitas amostras são registradas.

Dentre as diversas atividades de supervisão executadas no ambiente de tempo real utilizando este banco de dados histórico, aqui o interesse se revela para as inspeções periódicas ou aperiódicas (quando ocorre algum evento) realizadas na ferramenta de visualização gráfica do histórico de grandezas medidas pelo SCADA, no intuito de identificar possíveis anomalias, com base nas propriedades da grandeza que se deseja analisar. Por exemplo, a pressão do sistema de óleo do regulador de velocidade, usado para o comando do distribuidor da turbina que, por sua vez, controla a velocidade do conjunto turbina-gerador, tem comportamento repetitivo ao longo do tempo e mudanças neste comportamento podem indicar anomalias[1].

Observa-se que esta atividade de inspeção limita-se à capacidade humana de avaliar certa quantidade de dados relativos a um intervalo de tempo. Considerando a quantidade de grandezas que podem ser analisadas, se torna interessante utilizar ferramentas computacionais para processar e potencializar a capacidade de análise humana. A Figura 2.6 mostra um técnico da Sala de Controle de Despacho de Carga de Itaipu durante trabalho de supervisão de algumas telas do sistema SCADA e histórico.

Tradicionalmente, existem dezenas de milhares de grandezas historiadas pelo sistema de supervisão em grandes usinas hidrelétricas e, de maneira geral, apenas uma pequena parte destas grandezas é usada para algum tipo de análise mais elaborada do que o simples monitoramento de seus limites. Portanto, este trabalho propõe o uso de técnicas e ferramentas de mineração de dados para a exploração e contínua análise de dados selecionados.

2.3 Inspeções Físicas da Operação

Outra atividade realizada rotineiramente em usinas hidrelétricas que pode ser beneficiada com a pesquisa realizada neste trabalho são as Inspeções da Operação. Estas atividades são realizadas por operadores treinados, que periodicamente inspecionam fisi-

[1]Este sistema será detalhado no estudos de caso apresentados nas Seções 5.5 e 5.8

Figura 2.6: Técnico durante trabalho de supervisão.

camente equipamentos espalhados por toda a usina. Durante esta atividade de inspeção, são verificadas as condições gerais de funcionamento do equipamento, como ruído, temperatura, umidade, presença de vazamentos, bem como a leitura de instrumentos que não são medidos remotamente pelo sistema SCADA.

Esta atividade pode se beneficiar de indicações, provenientes de sistemas computacionais (tal como o aqui proposto), de possíveis comportamentos anômalos, que ainda não tenham provocado falha propriamente dita, mas que estejam na iminência de fazê-lo se não forem tomadas ações preventivas. Assim, com sugestões de possíveis anomalias, os operadores podem executar verificações adicionais ao inspecionar os equipamentos indicados para se certificar de que esteja tudo funcionando adequadamente.

Capítulo 3

Fundamentação Teórica

Neste capítulo, serão apresentados os principais conceitos e métodos encontrados na literatura relacionados à mineração de dados, séries temporais e detecção de anomalias, a serem aplicados neste trabalho.

3.1 Descoberta de Conhecimento em Dados

A atividade de mineração de dados busca extrair informações úteis para a tomada de decisão através da aplicação de técnicas automáticas em bases de dados existentes. Trata-se de atividade de real valor, pelos benefícios que a informação minerada pode trazer às instituições, reduzindo custos e aumentando a eficiência de processos em curso [12].

Importante ressaltar que, apesar da execução do algoritmo de mineração propriamente dito se tratar de um processo automático, esta atividade está inserida em um contexto maior, conhecido como *Knowledge Discovery from Data – KDD* [8], em que uma série de outros processos, nem sempre automáticos, são necessários para "preparar" a informação para ser minerada. Neste contexto, o termo "mineração" remete a uma árdua atividade de extração de informações realmente valiosas.

Adotando o processo de *KDD* como uma metodologia para a análise de dados, inicialmente, procede-se à seleção, pré-processamento e transformação dos dados. Uma das maneiras de executar estas tarefas, denominada *ETL* (*Extration, Transformation and Load*), faz com que os dados de entrada sejam extraídos de suas diversas fontes originais, limpos, integrados, transformados e, por fim, carregados em um único banco de dados, conhecido como *Data Warehouse*, que reúne todas as informações que serão futuramente processadas pelos algoritmos de mineração de dados [12].

Além da possibilidade de aplicação de algoritmos automáticos para a descoberta de conhecimento, um *Data Warehouse* permite ainda a execução de análises *online* por especialistas de seu conteúdo, conhecida como *OLAP* (*Online Analytical Processing*). Esta forma de análise é realizada através de ferramentas específicas, que permitem a exploração dos dados de forma multidimensional em busca de correlação entre as diversas dimensões geradas durante o processo de *ETL* [12].

Tecnicamente, todas estas formas de análise são possíveis porque um *Data Warehouse* tem a característica de ser fisicamente segregado dos bancos operacionais, evitando que possíveis sobrecargas no servidor, proveniente de consultas complexas, interfiram nos processos operacionais em curso [12].

Posteriormente, existe a etapa de execução dos algoritmos de mineração de dados propriamente ditos, que serão detalhados nas próximas seções. Nesta etapa são produzidos modelos que podem classificar, associar ou agrupar os dados resultantes da etapa anterior, agregando valor ao processo. Estes resultados podem ser também armazenados para avaliação e consulta futura.

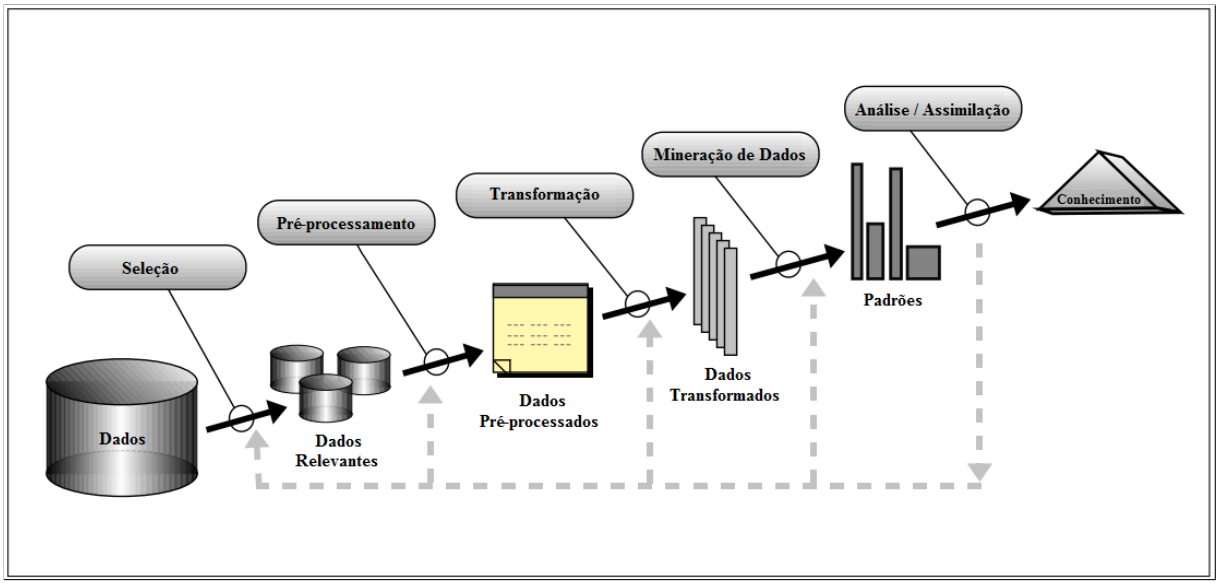

Figura 3.1: Etapas do processo de descoberta do conhecimento em dados [8].

3.2 Séries Temporais

Séries temporais são conjuntos de dados ordenados em função do tempo. Cada registro é chamado de amostra e está associado a certa marca temporal (conhecida como estampa de tempo). Normalmente são utilizadas para representar grandezas físicas, permitindo

reconstruir seu comportamento dinâmico ao longo do tempo.

Do ponto de vista de armazenamento computacional, um conjunto de séries temporais nada mais é do que uma tabela de um banco de dados com campos de identificação da série, estampa de tempo (*timestamp*) e valor e, portanto, podem ser usados como fonte para a construção de bancos de dados derivados, como descrito na seção anterior no processo de descoberta de conhecimento. Para isso, é preciso definir quais dados seriam importados e as funções de limpeza e transformação que seriam aplicadas às séries selecionadas. Estes dados seriam então movidos para um banco de dados (DW) e a estampa de tempo de cada amostra seria transformada em um atributo, que relacionaria todos os eventos que aconteceram em mesmo instante ou intervalo de tempo.

No contexto analisado neste trabalho, as séries temporais podem assumir dois tipos básicos distintos de valores. Séries *analógicas*, que representam valores reais contínuos no tempo e proporcionais a grandezas físicas, como temperatura e corrente elétrica, e séries de estado, que representam diferentes estados de um equipamento ou subsistema no tempo, como *ligado* e *desligado*. Sua diferenciação é importante para o pré-processamento dos atributos e aplicação dos algoritmos, pois as séries analógicas devem ser tratadas considerando a continuidade de sua magnitude e sinal ao longo do tempo, enquanto que séries de estados, apesar de serem representadas por valores numéricos discretos, não possuem qualquer associação do estado representado com o valor do número que o representa. Ou seja, considerando uma série temporal de estado, representada por valores 1, 2 e 3, o estado 1 desta série não tem uma relação mais próxima com o estado 2 do que com o estado 3, por exemplo. São apenas símbolos para representar diferentes estados.

Importante destacar também que bancos de dados de séries temporais em plantas industriais utilizam diferentes frequências de amostragem das grandezas de interesse. Assim, não é factível simplesmente selecionar uma amostra de uma dada grandeza em um determinado intervalo de tempo e tentar relacioná-la com a amostra de outra grandeza, pois pode ocorrer que tal amostra não exista no instante selecionado. Por esta razão, neste trabalho se emprega uma etapa de pré-processamento em que se realiza uma interpolação de dados para permitir a sincronização, ainda que em intervalos irregulares, de todas as grandezas envolvidas em uma determinada análise. Na Figura 3.2, pode-se observar que as três séries representadas (referentes às grandezas nível de óleo, temperatura e potência de uma unidade geradora) têm frequências de amostragem diferentes/irregulares.

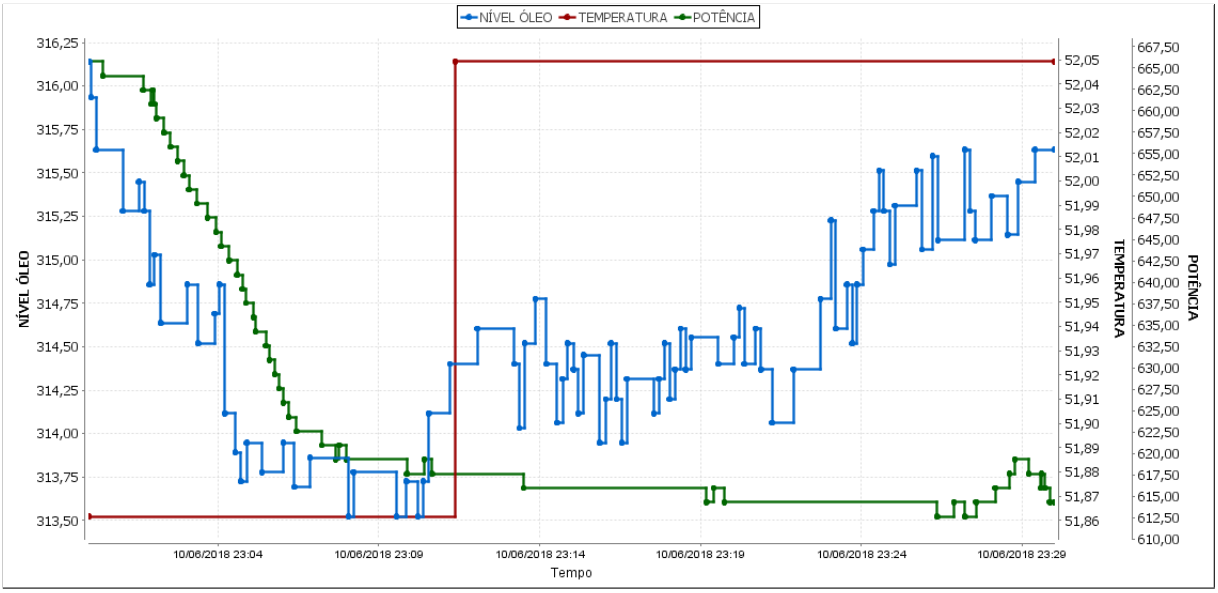

Figura 3.2: Gráfico de séries temporais com diferentes frequências de amostragem.

3.3 Detecção de *Anomalias* pela Mineração de Dados

Uma anomalia pode ser representada por uma amostra ou conjunto de amostras de uma série temporal que signifique um comportamento diferente do esperado ou normal [12, 19]. Usualmente, sua origem difere de ruídos no processo de aquisição dos dados e representa alterações físicas no processo que está sendo observado [10]. Diferentemente dos ruídos, inerentes a qualquer processo de observação e normalmente indesejados, as anomalias representam informações valiosas do processo, pois podem identificar um possível início de falha em um equipamento [15] e sua detecção está entre os principais interesses desta dissertação.

Com relação ao escopo, anomalias podem ser classificadas como globais ou contextuais. As anomalias globais são caracterizadas por se manifestarem em relação a todo o conjunto de amostras. Já as anomalias contextuais, se manifestam em relação a um conjunto de amostras específico (contexto) e podem não ser consideradas anomalias em outro conjunto de amostras (outro contexto). Ou seja, em muitos casos, os modelos que definem o comportamento de uma série temporal estão contextualizados às condições ou estados relativos a esta série [21]. Em um caso clássico, cita-se a temperatura observada nas estações do ano, como mostra a Figura 3.3, onde as amostras devem ser avaliadas dentro do contexto adequado (estação do ano) para poderem ser consideradas anômalas ou não.

Quanto à coletividade, anomalias podem ser classificadas como: pontuais, representadas por anomalias provocadas por uma única amostra no conjunto de dados; coletivas,

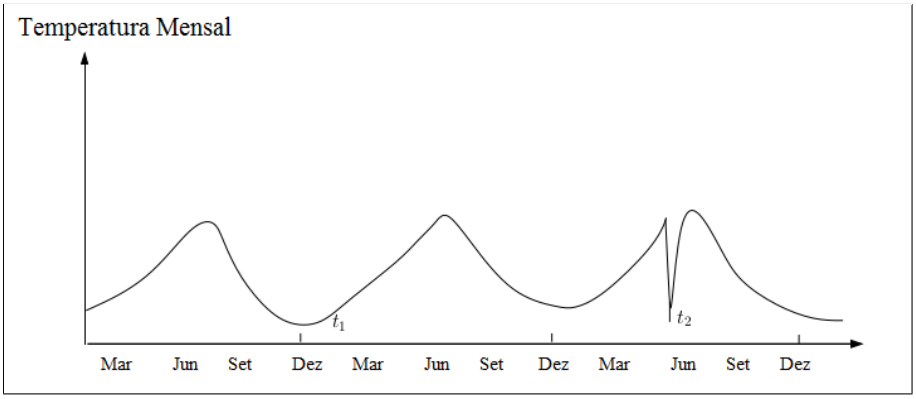

Figura 3.3: Temperaturas anômalas observadas em contexto anual [5].

aquelas formadas por um conjunto de amostras que juntas representam um comportamento anômalo. Nestes dois casos, as anomalias podem se manifestar em um escopo global ou contextual [12, 3].

Devido a impossibilidade de se considerar que todas as anomalias sejam sempre formadas por apenas uma amostra ou, ao contrário, sempre por um conjunto de amostras, será usado o termo *evento* para caracterizar uma ou mais amostras, subsequentes no tempo, que podem compor uma anomalia ao longo deste trabalho [3]. O termo *evento* tem ainda a vantagem de ser também familiar aos estudiosos do sistema elétrico, sendo usado para caracterizar momentos importantes que podem, por exemplo, resultar em perturbações.

3.3.1 Alteração do Contexto

Alteração do contexto ou *Concept Drift* (termo usado em inglês) é um fenômeno comum em dados de séries temporais resultantes de medições em sensores de processos físicos, como temperaturas ou pressões em equipamentos industriais, por exemplo. Consiste em uma significativa alteração no padrão dos dados, gradual ou abrupta, em um determinado contexto, sem contudo representar uma anomalia, não devendo portanto ser classificada como tal [23]. Este fenômeno se torna particularmente importante em sistemas que classificam novos dados através de um fluxo contínuo de amostras (*data streaming*) e podem implicar em falsos positivos e na necessidade de retreinar modelos.

Técnicas como o descarte de amostras muito antigas ou a redução no peso destas amostras no processo de decisão foram propostas em uma heurística que lembra o processo natural de esquecimento dos seres humanos [23, 17].

3.3.2 Rotulagem dos dados

Os algoritmos de mineração de dados também podem ser classificados de acordo com a existência ou não de rotulagem, que consiste em uma indicação para cada instância se ela se trata ou não de uma anomalia; os métodos supervisionados utilizam todos os dados rotulados por especialistas para treinar modelos com base na associação de regras previamente apresentadas.

Métodos semi-supervisionados utilizam apenas um subconjunto dos dados rotulados como, por exemplo, somente os dados normais.

Por fim, métodos não-supervisionados utilizam dados sem nenhuma rotulagem (informação) a respeito da sua classificação, o que é muito comum no mundo real. Para funcionarem adequadamente, muitas abordagens assumem que os dados normais são muito mais frequentes que as anomalias. Alguns algoritmos então se dedicam a aprender o que é normal no conjunto [3, 12].

3.3.3 Desbalanceamento entre Classes

Métodos tradicionais de classificação assumem que os dados estão uniformemente distribuídos entre as diferentes classes existentes. Porém, isso muitas vezes não é verdadeiro no mundo real, onde encontram-se situações em que a classe de interesse representa apenas um pequeno percentual de instâncias do conjunto de treinamento. Este tipo de situação, denominado *desbalanceamento entre classes*, pode acarretar em um resultado insatisfatório do processo de treinamento e, consequentemente, do resultado final do classificador. Por isso, diversos métodos para minimizar este problema foram propostos. Neste trabalho, será usada uma técnica simples de sobre-amostragem, em que novas instâncias da classe sub-representada são geradas para minimizar os efeitos deste problema. Esta técnica, chamada de SMOTE (synthetic minority over-sampling technique), gera sinteticamente novas instâncias para classes minoritárias baseadas nos seus vizinhos próximos [4].

Em tempo, alguns algoritmos não supervisionados para detecção de anomalias utilizam justamente a premissa de que a classe que representa instâncias anômalas ocorre em uma frequência muito menor do que as instâncias de classe normal. Nestes casos, a utilização de métodos para balancear as classes prejudica o resultado do algoritmo.

3.3.4 Métodos para Detecção de Anomalia

A literatura apresenta diversas categorias de métodos para a detecção de anomalias, descritas brevemente a seguir.

Métodos estatísticos são aqueles que assumem que os dados seguirão uma distribuição baseada em um modelo estatístico. Estes métodos podem ser divididos em modelos pré-definidos, em que os dados seguem uma distribuição conhecida, e modelos que serão identificados por meio da análise dos dados (através do uso de histogramas, por exemplo). Quando um objeto se desviar significativamente do comportamento dele esperado por seu modelo, considera-se presente uma anomalia.

Métodos baseados em análise de proximidade avaliam se a distância entre um evento e seus vizinhos mais próximos se desvia significativamente da distância entre seus eventos vizinhos e os vizinhos deles. Existe ainda a análise da densidade entre um evento e seus vizinhos, em que um evento é considerado como uma anomalia quando sua densidade se desvia significativamente de seus vizinhos. Por fim, métodos baseados em agrupamento (*clustering-based*) são aqueles em que anomalias são encontradas em grupos pequenos e de baixa densidade.

A detecção de anomalias também pode ser considerada um subgrupo da tarefa de classificação que tem o propósito de criar, a partir de dados históricos, modelos para classificação de apenas uma classe (*one-class model*) para identificação de objetos de classe normal. Os objetos não classificados como pertencentes a esta classe são considerados anomalias. Os modelos de classificação podem ser categorizados ainda de acordo com o momento em que são construídos; se construídos antes da classificação de qualquer instância são chamados de *eager*; já os avaliados no momento da classificação de uma instância são chamados de *lazy* [3, 12].

3.4 Algoritmos de Mineração de Dados

A seguir, são descritos os algoritmos de mineração de dados aplicados neste trabalho de pesquisa: dois não-supervisionados e dois supervisionados, explorados no Capítulo 4, com resultados e características comparadas no Capítulo 5.

3.4.1 Rede Neural Replicante

Método *eager*, não supervisionado, tem origem em modelos utilizados para compressão de dados de imagens. Uma rede neural replicante (RNN) utiliza uma rede *Multi-Layer Perceptron* com o objetivo de reproduzir (replicar) os dados de entrada na saída [14]. Cada neurônio é do tipo *Perceptron* e funciona através de modelos matemáticos que são inspirados nos mecanismos de funcionamento de um neurônio biológico. Estes neurônios são compostos por um conjunto de entradas e uma saída. Cada saída é conectada às entradas dos neurônios da camada seguinte através de pesos, que são calculados durante o treinamento. Cada neurônio realiza a soma das suas entradas, ponderadas pelos respectivos pesos. Este valor é somado a um *bias* e aplicado a uma função de ativação que determina a sua saída que, por sua vez, representa uma das entradas de cada neurônio da camada seguinte I_j, definida pela equação 3.1.

$$I_j = \sum_{i}^{n} w_{ij} O_i + \theta_j \tag{3.1}$$

Onde w_{ij} é o peso da conexão do neurônio i da camada anterior para o neurônio j. O_i é a saída do neurônio i da camada anterior e θ_j é o *bias* do neurônio j da camada atual. Este *bias* serve como um *threshold* para a ativação do neurônio.

A saída de cada neurônio O_j é calculada aplicando o somatório das entradas da camada anterior I_j na função de ativação, conforme Equação 3.2. A função de ativação usada neste trabalho será a função *sigmóide* ou *logística*.

$$O_j = \frac{1}{1 + e^{-I_j}} \tag{3.2}$$

O treinamento da rede será feito pelo tradicional método *backpropagation*, que consiste em propagar o erro das saídas de cada camada a partir da última camada até a primeira, atualizando os *pesos* e *bias*, conforme descrito por Han et al [12]. Assim, para um neurônio j na camada de saída, considerando a função de ativação *sigmóide*, o erro pode ser calculado como:

$$Erro_j = O_j(1 - O_j)(T_j - O_j), \tag{3.3}$$

Onde O_j é a saída atual e T_j é o valor correto conhecido para esta saída que, no caso de uma RNN, é uma cópia da entrada. $O_j(1 - O_j)$ é a derivada da função *sigmóide*.

Para computar os erros da camada anterior (oculta), o somatório ponderado dos erros da camada seguinte é considerado. Portanto, o erro do neurônio j de uma camada oculta pode ser calculado como:

$$Erro_j = O_j(1 - O_j) \sum_k Erro_k w_{jk} \tag{3.4}$$

Onde w_jk é o peso da conexão entre o neurônio j e o neurônio k da próxima camada. Os pesos e bias são então atualizados proporcionalmente ao erro propagado entre as camadas através das equações:

$$w_{ij} = w_{ij} + l.Erro_j O_i \tag{3.5}$$

$$\theta j = \theta j + l.Erro_j \tag{3.6}$$

Onde l é a taxa de aprendizagem do modelo.

Este processo é executado iterativamente e cada ciclo completo de atualização é chamado de *época*. O processo de treinamento é encerrado quando o erro for menor do que um limite especificado ou o número de épocas for maior do que o máximo permitido.

A particularidade da rede neural replicante é ser formada pelo mesmo número de neurônios na camada de entrada e de saída para ser treinada, como foi mencionado, para reproduzir a entrada na saída. Ou seja, ela é treinada usando como rótulos a própria entrada. Além disso, a camada oculta contém um número menor de neurônios, provocando o efeito de compressão dos dados, já que para ter os mesmos dados da entrada na saída, será preciso representá-los, de alguma forma, com menos neurônios nas camadas ocultas, como mostra a Figura 3.4. Desta forma, ao treinar esta rede e considerando que a grande maioria dos dados são normais, espera-se que ela *aprenda* a reproduzir melhor estes tipos de dados. Assim, o erro (distância) entre os valores de entrada e a saída prevista representa a raridade dos eventos e, por conseguinte, a probabilidade de se tratar de uma anomalia. Assim, em aplicações práticas, é necessário definir um limite (*threshold*) para esta distância para que uma instância seja considerada objetivamente uma anomalia.

Em 2002, Hawkins [14] propôs uma rede com cinco camadas, três ocultas, uma de entrada e outra de saída. Esta configuração foi escolhida empiricamente com o objetivo de minimizar os erros durante o treinamento. Em 2014, Dau et al. [7] demonstrou que a utilização de apenas uma camada oculta teria bons resultados e simplificaria muito o processo, especialmente a definição dos parâmetros da rede. No aplicativo desenvolvido

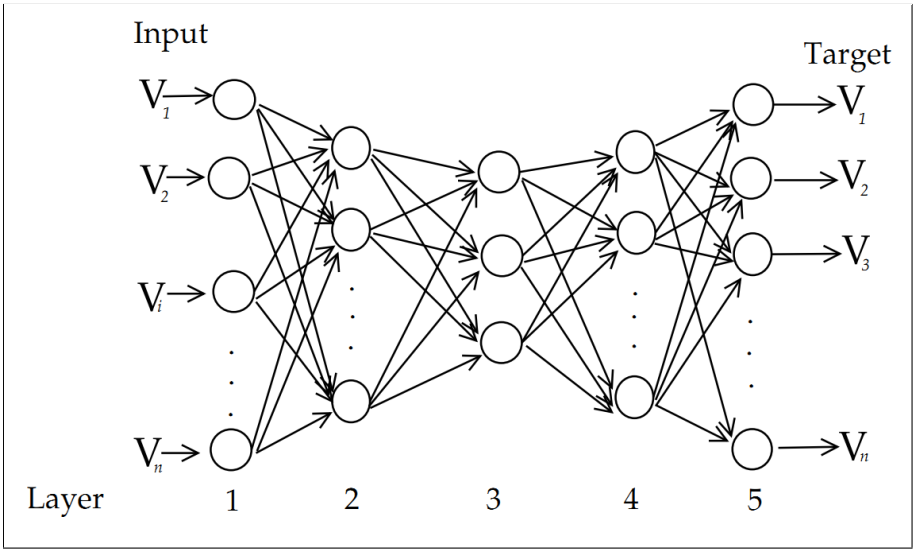

Figura 3.4: Exemplo de rede neural com treinamento não-supervisionado para detecção de anomalia [14]

para este trabalho, foi utilizada uma rede neural com 3 camadas, com o mesmo número variável de neurônios na camada de entrada e saída, definido de acordo com o número de atributos selecionados pelo usuário na etapa de configuração, e uma camada oculta com um número menor de neurônios que as camadas de entrada e saída.

3.4.2 Árvore de Decisão J48

O algoritmo J48 consiste em um algoritmo supervisionado que cria uma árvore de decisão, baseada na relevância de cada atributo para determinar a classe da instância. Este é um algoritmo *eager*, pois cria-se uma árvore com os dados de treinamento e, posteriormente, sua execução consiste em apenas aplicar um novo atributo ao modelo já criado. Ele foi baseado no algoritmo *C4.5*, que por sua vez é uma versão modificada do algoritmo *ID3*, que utiliza o conceito de entropia para construir a referida árvore [20].

Árvores de decisão são estruturas similares a fluxogramas em formato de árvores, em que: cada nó intermediário representa um teste em um atributo; suas saídas são os ramos da árvore; e as folhas, ou nós terminais, representam as classes de saída do algoritmo para uma determinada entrada.

A construção de uma árvore de decisão é feita através da aplicação de um algoritmo que cria nós e particiona as instâncias em subconjuntos recursivamente, a partir de um único nó até as folhas da árvore. Dado um nó qualquer, quando todas as instâncias da partição resultantes a este nó são da mesma classe, então ele é definido como uma folha

que retorna a classe de suas instâncias como resultado da árvore. Senão, é verificado se a lista de atributos ainda não avaliados está vazia. Se sim, o nó também será definido como uma folha que retornará a classe majoritária de suas instâncias como resultado. Por fim, caso o nó atual não seja resolvido como uma folha, é aplicado um processo de seleção de outro atributo e criado um novo nó. Este processo é realizado recursivamente até que sejam definidas todas as folhas.

O critério de seleção de atributos é o núcleo do algoritmo de construção das árvores de decisão, sendo formado por uma heurística que tenta encontrar o atributo que melhor separa as instâncias em determinado nível da árvore. O algoritmo *ID3* usa um critério chamado de *Ganho de Informação* para a seleção. Este critério consiste em determinar qual atributo minimiza a informação necessária para classificar as instâncias nas partições resultantes. A informação esperada necessária para classificar uma instância em D obtém-se por:

$$Info(D) = -\sum_{i=1}^{m} p_i log_2(p_i), \tag{3.7}$$

Sendo p_i a probabilidade de uma instância em D pertencer a classe C_i.

A informação necessária para classificar uma instância em uma partição resultante da seleção do atributo A é definida como:

$$Info_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} Info(D_j), \tag{3.8}$$

Sendo D_j a partição resultante da separação usando o atributo A e o termo $\frac{|D_j|}{|D|}$ serve como um peso para cada partição, resultando em um valor ponderado. Por fim, o ganho de informação da seleção de cada atributo é definido como a diferença entre o ganho de informação da partição original e cada uma das novas partições possíveis. Assim, o ganho de informação resultante da seleção do atributo A pode ser definido como:

$$G(A) = Info(D) - Info_A(D) \tag{3.9}$$

Desta forma, o atributo que prover o maior ganho de informação e, por conseguinte, minimizar a informação necessária para classificar as instâncias das partições resultantes, será o escolhido.

3.4.3 Isolation Forest

Método *eager* não supervisionado, derivado de princípios do método supervisionado de árvore de decisão, apresenta-se como muito eficiente para a detecção de anomalias. O método baseia-se na ideia de que se for construída uma árvore que separe cada amostra de um conjunto de dados, então as anomalias serão isoladas em uma profundidade muito menor, ou seja, chegarão à folha da árvore muito mais rapidamente do que os valores normais. Isto ocorre porque elas se apresentam em uma quantidade muito menor e o valor de seus atributos difere muito dos demais, facilitando a sua separação [18].

A detecção de anomalias usando *Isolation Forest* ocorre em dois estágios. No primeiro, chamado de treinamento, são construídas árvores de isolação usando subconjuntos do conjunto completo de treinamento. No segundo estágio as instâncias de teste são aplicadas às árvores para se obter um *score* de anomalia para cada instância.

Na fase de treinamento, as árvores (*iTrees*) são construídas recursivamente particionando as instâncias até que todas estejam isoladas. Inicialmente, uma floresta de tamanho t é inicializada e o conjunto de treinamento completo X é particionado em subconjuntos X' de tamanho ψ, que serão usados em cada uma das árvores. Cada árvore da floresta é construída selecionando aleatoriamente atributos das instâncias contidas em seu subconjunto X'. O ponto de divisão de cada atributo também é selecionado aleatoriamente entre o valor mínimo e máximo do atributo selecionado. No final deste processo, uma coleção de árvores é retornada e está disponível para classificação.

Na fase de execução, calcula-se o tamanho do caminho percorrido por uma instância para ser classificada. Assim, o algoritmo conta recursivamente o número de nós percorridos por uma instância até chegar a uma folha, limitando este valor a um tamanho máximo definido por *hlim*. É justamente este tamanho do caminho percorrido por uma instância na árvore que serve como indicativo de anomalia.

Por fim, um *score* normalizado é calculado, pois as árvores têm estruturas diferentes, não sendo possível comparar diretamente o tamanho dos caminhos produzidos por elas. Para isso, o tamanho médio dos caminhos é calculado através de um método originário da busca em árvores binárias, conforme as regras definidas a seguir:

$$c(\psi) = \begin{cases} 2H(\psi - 1) - 2(\psi - 1) & para\ \psi > 2, \\ 1 & para\ \psi = 2, \\ 0 & para\ \psi < 2, \end{cases} \tag{3.10}$$

Sendo $H(i)$ o número harmônico, estimado por $ln(i) + 0,5772156649$ (constante de Euler). Considerando que $c(\psi)$ é a média do tamanho dos caminhos $h(x)$ dado ψ, ele pode ser usado para normalizar $h(x)$. Assim, o *score* normalizado de uma anomalia pode ser definido como:

$$s(x, \psi) = 2^{-\frac{E(h(x))}{c(\psi)}} \tag{3.11}$$

Onde $E(h(x))$ é a média de $h(x)$ de uma coleção de $iTrees$.

De acordo com Liu et al. (2008) [18], valores de s próximos de 1 indicam anomalia. Se os valores de s forem muito menores que $0,5$ são provavelmente instâncias normais e, se todos os valores de s estiverem próximos de $0,5$, o conjunto não tem anomalias distintas.

3.4.4 K-Nearest Neighbors

O k-Nearest Neighbors (k-NN) é um algoritmo supervisionado que determina a classe de uma nova instância com base na classe das instâncias mais próximas (vizinhas). Permite a parametrização de quantas instâncias serão utilizadas para determinar a classe resultante e qual a forma de cálculo da distância que será avaliada. É considerado um classificador *Lazy*, pois faz a avaliação no momento da classificação, ou seja, calcula a distância dos elementos e faz a busca pelos mais próximos em tempo de execução [6]. Neste trabalho foi utilizada a distância euclidiana para definir a medida de proximidade entre duas instâncias (x_1 e x_2) com n atributos, definida por:

$$Dist = \sqrt{\sum_{i=1}^{n} (x_{1i} - x_{2i})^2} \tag{3.12}$$

Capítulo 4

Metodologia Proposta para Mineração de Dados Operativos

4.1 Introdução

Neste capítulo, propõe-se uma metodologia para a utilização de técnicas de mineração de dados operativos para tratar diversos tipos de problemas existentes em usinas hidrelétricas, que serão apresentados no Capítulo 5. Esta metodologia é destinada a promover uma interação equilibrada entre o conhecimento de especialistas do setor elétrico de potência e as técnicas apresentadas no Capítulo 3.

No decorrer da presente pesquisa e desenvolvimento, observou-se que o processo de descoberta de conhecimento se apresenta como adequado para tratar os problemas em estudo. Desta forma, a metodologia proposta comporta três etapas relacionadas ao *KDD*, a saber: seleção e transformação de dados; construção do conjunto de treinamento supervisionado; e execução periódica para fins de diagnóstico.

Importante observar que o desenvolvimento da metodologia foi feito em conjunto com o desenvolvimento de uma ferramenta computacional que a implementa. Isto permitiu avaliar seu uso pelos especialistas e propor mudanças para adequá-la à realidade do seu setor de aplicação, em um processo iterativo. Por isso, a apresentação desta metodologia, neste capítulo, usa exemplos concretos, presentes na ferramenta computacional, para clarificar seu entendimento.

4.2 Características do Problema e Estratégias Adotadas

Inicialmente, deve-se definir as principais características dos problemas a serem analisados e as estratégias adotadas para tratá-los, como será visto a seguir.

4.2.1 Contextualização dos Dados

Os problemas abordados podem ser classificados como aqueles de identificação de anomalias contextuais, i.e. aquelas que se manifestam em certas condições de operação, podendo não ser anômalas em condições de operação diversas. No caso das grandezas de uma planta de produção de energia elétrica, o contexto de cada série pode ser determinado por estados operativos associados a esta série. Ou seja, a informação dos possíveis estados operativos de uma unidade geradora pode ser utilizada para determinar o número de *clusters* que uma grandeza associada a esta unidade geradora pode ter. Desta forma, espera-se que uma unidade geradora que esteja sincronizada no sistema interligado[1] tenha uma medida de frequência muito próxima de 60Hz, por exemplo.

Além disto, outra possível forma de contextualizar um conjunto de dados considera a hipótese de que, devido ao porte da usina (no caso de Itaipu, composta de 20 unidades geradoras de aproximadamente 700 MW cada), o comportamento de uma grandeza de certa unidade deve ser muito parecido ao comportamento da mesma grandeza das outras unidades que estejam no mesmo estado operativo. Assim, é possível correlacionar as medidas entre as unidades, em um mesmo instante ou intervalo de tempo, de modo a criar um padrão considerado normal, assumindo que a maioria das unidades que estão operando nestas condições têm um comportamento normal, mesmo sem definir formalmente os parâmetros desta "normalidade". Um exemplo desta condição pode ser definido como o desvio de potência ativa de cada uma das unidades em relação à média das outras unidades sincronizadas na rede elétrica do sistema e em modo de controle AUTO no sistema de controle automático da geração (CAG). Em condições normais de operação, este valor tende a zero, mas uma série de anomalias provoca, de tempos em tempos, um aumento significativo deste valor.

[1] A unidade geradora sincronizada ao sistema interligado é um dos seus possíveis estados operativos.

4.2.2 Intervalo de Agregação

Os problemas também podem ser caracterizados como a identificação de anomalias coletivas, pois são representadas por um conjunto de amostras, como ilustra a Figura 4.1, onde é possível observar que a série *anômala* (acima) apresenta valores individuais (amostras) dentro da faixa de valores da série *normal* (abaixo). Assim, não é possível identificar uma anomalia processando as amostras individualmente. Para lidar com esta característica, será utilizada uma estratégia de redução do conjunto de amostras em atributos específicos, transformando assim o problema em um caso de detecção de anomalia pontual, com diversos atributos que agregam o conjunto de amostras [9]. Porém é importante observar que esta agregação pode resultar em perda substancial de informação, como ocorre quando os dados são agregados por média, por exemplo, que deixa de representar comportamentos importantes para a detecção de anomalias.

Além disso, a quantidade de dados a ser analisada, ou seja, o intervalo ou janela de tempo que deve ser considerada adequada para a agregação também é muito importante para a identificação de anomalias e representa um desafio para este trabalho. Isto porque existem grandezas, como a potência de uma unidade geradora, que não apresentam um comportamento constante como a frequência elétrica[2], e podem variar ao longo do tempo e permanecer na condição de normalidade. Desta forma, além da contextualização, deve-se definir o intervalo temporal que representa todo um ciclo que pode conter a anomalia coletiva, do início ao fim, para determinar se um conjunto de amostras realmente representa uma anomalia.

4.2.3 Desvio de Contexto

Além da contextualização e da coletividade, algumas séries temporais analisadas nesta pesquisa também apresentam desvio de contexto (ver Seção 3.3.1), quando observados em um grande período. Isto ocorre devido a alterações físicas nos equipamentos, que podem ter peças substituídas por outras de modelos diferentes após uma manutenção, e passarem a ter um comportamento diferente após voltarem a operar. Alterações na regulagem e o desgaste natural dos equipamentos também provocam mudanças ao longo da vida útil. Por isso, é importante tratar os efeitos desta mudança do contexto para que ela não se manifeste como falsos positivos no sistema.

[2] A frequência elétrica no sistema interligado nacional apresenta valores muito próximos a 60Hz em condições normais de operação.

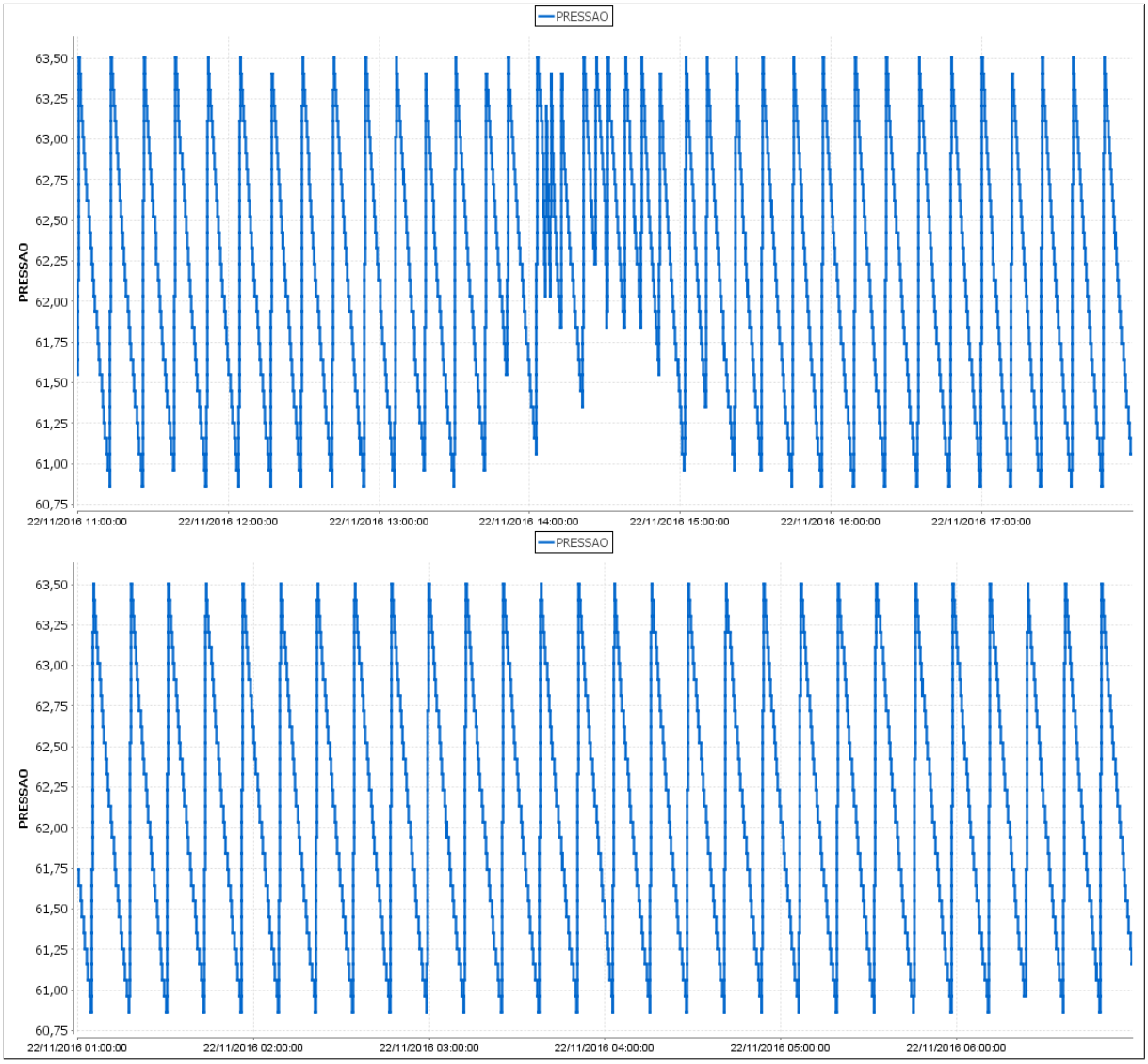

Figura 4.1: Série com comportamento *anômalo* acima e *normal* abaixo.

Na Figura 4.2, observa-se um desvio de contexto da pressão do tanque de óleo do regulador de velocidade, que será detalhado adiante no estudo de caso da Seção 5.8. Observa-se que esta figura apresenta um gráfico de 335 dias e, em determinado momento, é possível ver claramente uma mudança de patamar da grandeza. Esta mudança influenciará atributos como média, máximo e mínimo, o que deve ser analisado com cuidado.

Além disso, atributos como frequência e duração, que serão definidos na Seção 4.3.3, quando usam o desvio padrão como limite para definir as faixas de variação, também serão afetados. Por isso, o cálculo do desvio padrão global, que considera todo o período de análise e não apenas um único intervalo, deve considerar eliminar valores mais antigos, permitindo assim mudanças gradativas de contexto.

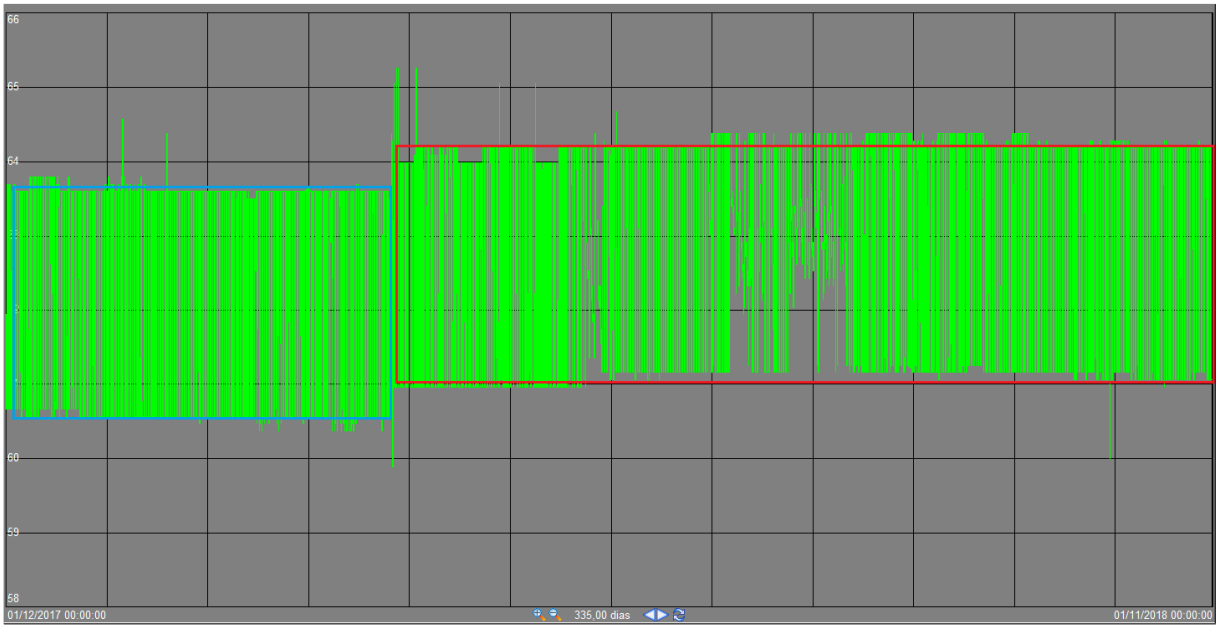

Figura 4.2: Exemplo de alteração de contexto, destacados em azul e vermelho.

4.2.4 Rotulagem de Dados

Apesar da grande quantidade de dados, não existe atualmente uma rotulagem dos dados armazenados no banco de dados histórico indicando anomalias e, tampouco correlacionando as séries de forma a encontrar atributos significativos para sua identificação. Ou seja, atualmente o banco de dados consiste em um conjunto de séries históricas isoladas e não-rotuladas e cabe ao especialista relacioná-las e identificar manualmente possíveis comportamentos anômalos. Portanto, considerando a ausência de rotulagem das informações, o problema se caracteriza por ser não-supervisionado.

Para minimizar a dificuldade de realizar manualmente a rotulagem de uma grande quantidade de dados, foi definida na metodologia uma etapa de pré-rotulagem dos dados, na qual será aplicada um classificador não supervisionado para gerar um conjunto de treinamento rotulado preliminar. O resultado desta classificação será revisado pelo especialista, que fará pequenos ajustes, apenas em casos que considerar inadequados, e então o aprovará para uso em métodos supervisionados.

4.3 Seleção e Transformação dos Dados

Após apresentar as estratégias que serão utilizadas para tratar os desafios encontrados neste trabalho, cabe agora abordar a primeira etapa da metodologia para mineração dos dados operativos. O objetivo desta etapa é selecionar e transformar um conjunto de

tamanho variável de séries temporais de diversas grandezas físicas. Estas estão irregularmente intervaladas no tempo, portanto devendo ser transformadas em um conjunto de registros estruturados com intervalos regulares e um número fixo de atributos, para que seus elementos possam ser comparados entre si e empregados em algoritmos de mineração de dados. Isto implica em agregar, de alguma maneira, este conjunto de dados no intervalo que foi definido para realizar a redução descrita na Seção 4.2.2. As próximas Seções apresentam como isto pode ser feito de forma concreta, utilizando a ferramenta computacional desenvolvida neste trabalho.

4.3.1 Seleção dos Dados

Nesta primeira etapa do processo de descoberta de conhecimento, os dados brutos serão selecionados pelo usuário especialista, denominado doravante simplesmente de usuário, através da interface gráfica do aplicativo desenvolvido nesta Dissertação exibida na Figura 4.3. Ele terá à disposição diversas fontes distintas de dados, como dados históricos do sistema SCADA e dados sobre intervenções nos equipamentos. Esta integração de distintas fontes foi possível através da criação de uma interface comum, chamada de *Time Series*, implementada para diversas fontes. Assim, todos os dados de entrada deste sistema têm o formato de séries temporais e podem ser tratados como tais pela aplicação. Para a inclusão de uma nova fonte de dados é preciso apenas implementar a interface de acesso à fonte.

Para a seleção das fontes de dados, o usuário deve clicar no botão "+", indicado em (1) na Figura 4.3, e em seguida informar o termo de busca (2). O termo informado será então pesquisado nas diversas fontes e apresentado ao usuário na forma de uma lista para seleção (3). O usuário pode selecionar quantas fontes quiser nesta etapa e cabe a ele identificar séries (pontos de supervisão) que podem estar correlacionadas a anomalias.

4.3.2 Pré-processamento e Transformação

A seguir, o usuário deve definir uma equação que irá transformar os dados selecionados, usando para isto a aplicação de uma fórmula simples, indicada em (2) na Figura 4.4, com operadores matemáticos básicos, algumas funções lógicas disponíveis e as variáveis selecionadas na etapa de seleção dos dados. Considerando o intervalo irregular de amostragem das grandezas, os dados serão interpolados para todos os instantes de tempo existentes, em ao menos uma das grandezas, antes da aplicação da equação.

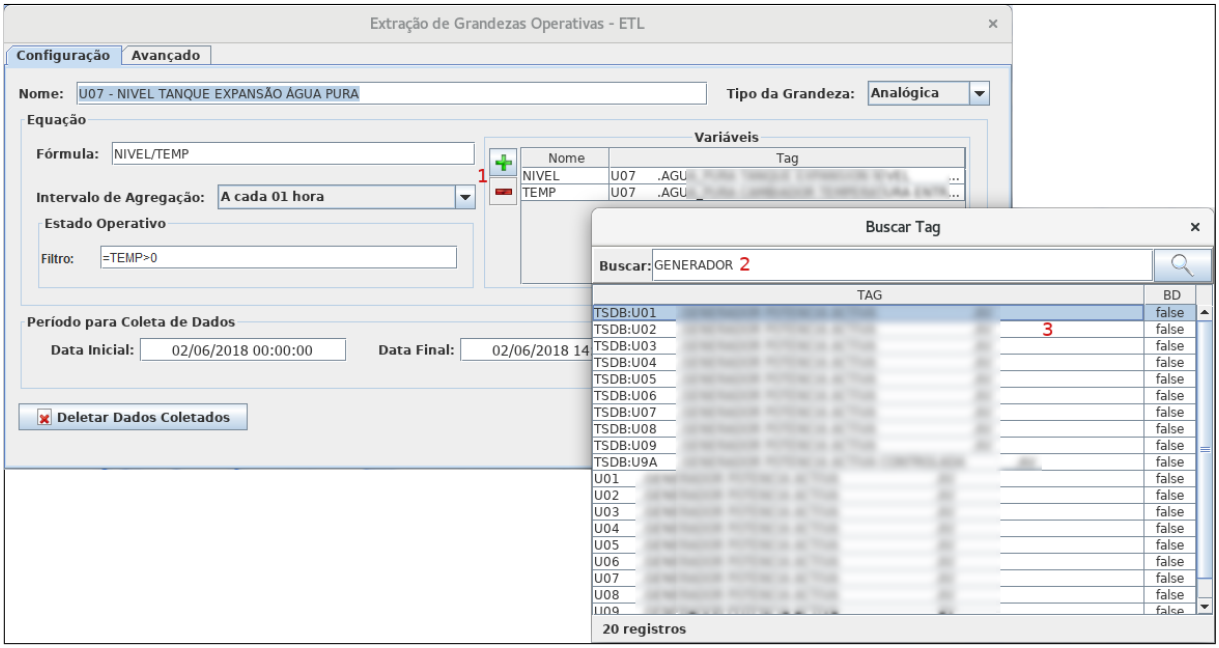

Figura 4.3: Interface para a seleção de fontes de dados.

O usuário também será responsável por definir o intervalo (janela) de tempo na qual serão aplicadas as funções de agregação dos dados, indicado em (3) na Figura 4.4. Estas funções serão compostas basicamente por propriedades estatísticas do conjunto de dados existente no intervalo, como média, desvio padrão, valores mínimo e máximo. Além disso, outras propriedades como frequência e duração (ver Seção 4.3.3) também serão calculadas. Cada uma destas propriedades será definida como um atributo relativo à grandeza selecionada.

Posteriormente, o usuário deve informar outra equação que irá definir o contexto da análise, chamado de estado operativo, indicado em (4) na Figura 4.4. Esta equação funcionará como um filtro e todos os dados referentes a estados operativos em que esta equação não retorne a condição de *verdadeiro* serão descartados.

Por fim, independentemente do tipo dos dados selecionados pelo usuário, pode-se escolher se o resultado da equação definida será processado como um valor *analógico* ou de *estado*, indicado em (5) na Figura 4.4, conforme propriedades definidas na Seção 3.2. A análise analógica consiste em considerar a saída os dados como valores contínuos no tempo e proporcionais a grandezas físicas. Assim, propriedades estatísticas como a média e o desvio padrão fazem sentido neste tipo de análise. Por outro lado, a análise de estados considera que os valores resultantes da equação consistem em valores discretos que representam estados do sistema, como *ativo* ou *inativo* e, portanto, desconsideram propriedades como a média ou o desvio padrão. Na análise de estado, outras propriedades

(e.g. moda), são utilizadas.

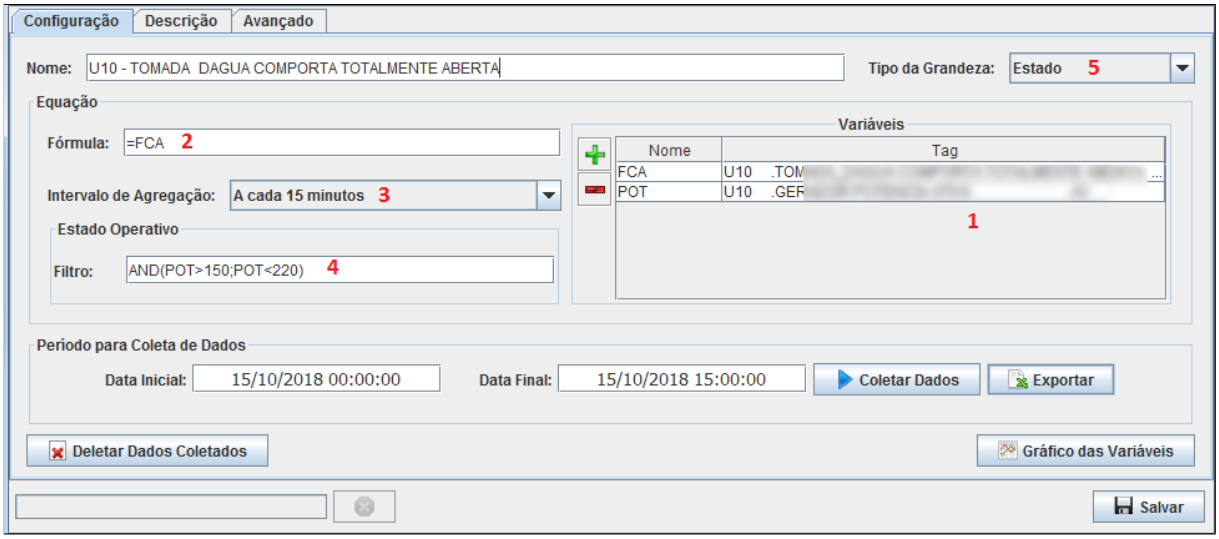

Figura 4.4: Interface gráfica de configuração do aplicativo de detecção de anomalias.

Após a configuração feita pelo usuário, os dados serão coletados das fontes selecionadas e processados conforme o período definido pelo usuário. Por fim, os resultados são escritos em um banco de dados, chamado de *Data Warehouse*, permitindo o seu uso para o treinamento de algoritmos de classificação, conforme as etapas seguintes.

Vale destacar que esta etapa terá o benefício adicional de disponibilizar um *Data Warehouse* para os especialistas realizarem análises OLAP, utilizando ferramentas de *Data Analytics*.

4.3.3 Atributos Agregados

Para fazer a redução da coletividade das amostras para eventos pontuais, que podem ou não representar anomalias e possuem um formato adequado para serem processados pelos algoritmos de mineração de dados, deve-se definir um conjunto de atributos agregados. Isto é feito para permitir a representação dos dados originais em intervalos regulares pré-definidos, conforme venha a ser configurado pelo usuário, sem que se percam suas propriedades originais que identificam uma anomalia.

Para isso, define-se um conjunto fixo de diferentes funções de agregação às quais todos os dados (resultantes da execução da equação descrita na Seção 4.3.2) são aplicados. Assim, o usuário poderá escolher, neste conjunto de atributos, quais os que devem ser usados como entrada para os modelos de classificação, de acordo com o caso específico que estiver analisando. A princípio, estas funções de agregação são qualquer transformação

aplicada ao conjunto de amostras, resultante da equação que relaciona as grandezas, que reduza este conjunto a apenas um número fixo de elementos no intervalo especificado.

Assim, para cada intervalo, são realizados cálculos estatísticos básicos (média, desvio padrão, valor máximo e valor mínimo) para o caso das análises analógicas e moda para análises de estado. Tais cálculos são feitos usando algoritmos de processamento de dados sequenciais (*streaming*) para que não seja necessário percorrer múltiplas vezes todo o conjunto de dados.

Além disso, é feita a contagem do número de amostras para o período para tirar proveito do efeito da compressão aplicada nos dados, descrito na Seção 2.2. De certa forma, a compressão funciona como um filtro para a grandeza e a variação brusca no número de amostras de um período, em relação aos demais, pode indicar um comportamento anormal da série ali representada.

Adicionalmente, em alguns casos, os dados no domínio do tempo podem não ser a melhor forma de representar o comportamento da uma grandeza para detectar anomalias. Neste sentido, transformações para outros domínios de representação podem facilitar muito este trabalho, como a transformação para o domínio da frequência de ativação, no caso da bomba do óleo do regulador de velocidade, que será apresentado na Seção 5.5, que permite facilmente identificar um vazamento no sistema.

Desta forma, dois outros atributos também calculados nesta etapa são *frequência* e *duração* em faixa discreta de dados. Para análises analógicas, estes atributos são obtidos após a definição de uma faixa discreta, definida como a média móvel global das amostras somada ao desvio padrão global[3]. Para análises de estado, a faixa é definida como qualquer estado diferente da moda de todas as amostras acumuladas desde o início do processamento até o instante presente. Após a definição da faixa, é contabilizado o tempo de permanência (duração) e a frequência de ativação das amostras do intervalo em processamento, conforme mostra a Figura 4.5.

Para o cálculo da média móvel global e seu respectivo desvio padrão, foi definido um algoritmo acumulador que descarta os valores mais antigos quando o desvio provocado pela adição de uma nova amostra (mais recente) ao conjunto é menor do que um limite, definido previamente e configurável pelo usuário. Assim, o valor da média (global), usada para o cálculo da frequência e duração, representa um conjunto de intervalos, o que é mais estável do que apenas um intervalo, mas também é capaz de se adaptar a uma nova condição permanente do sistema, definida como alteração do contexto na Seção 3.3.1,

[3]Calculadaos para todo o período analisado.

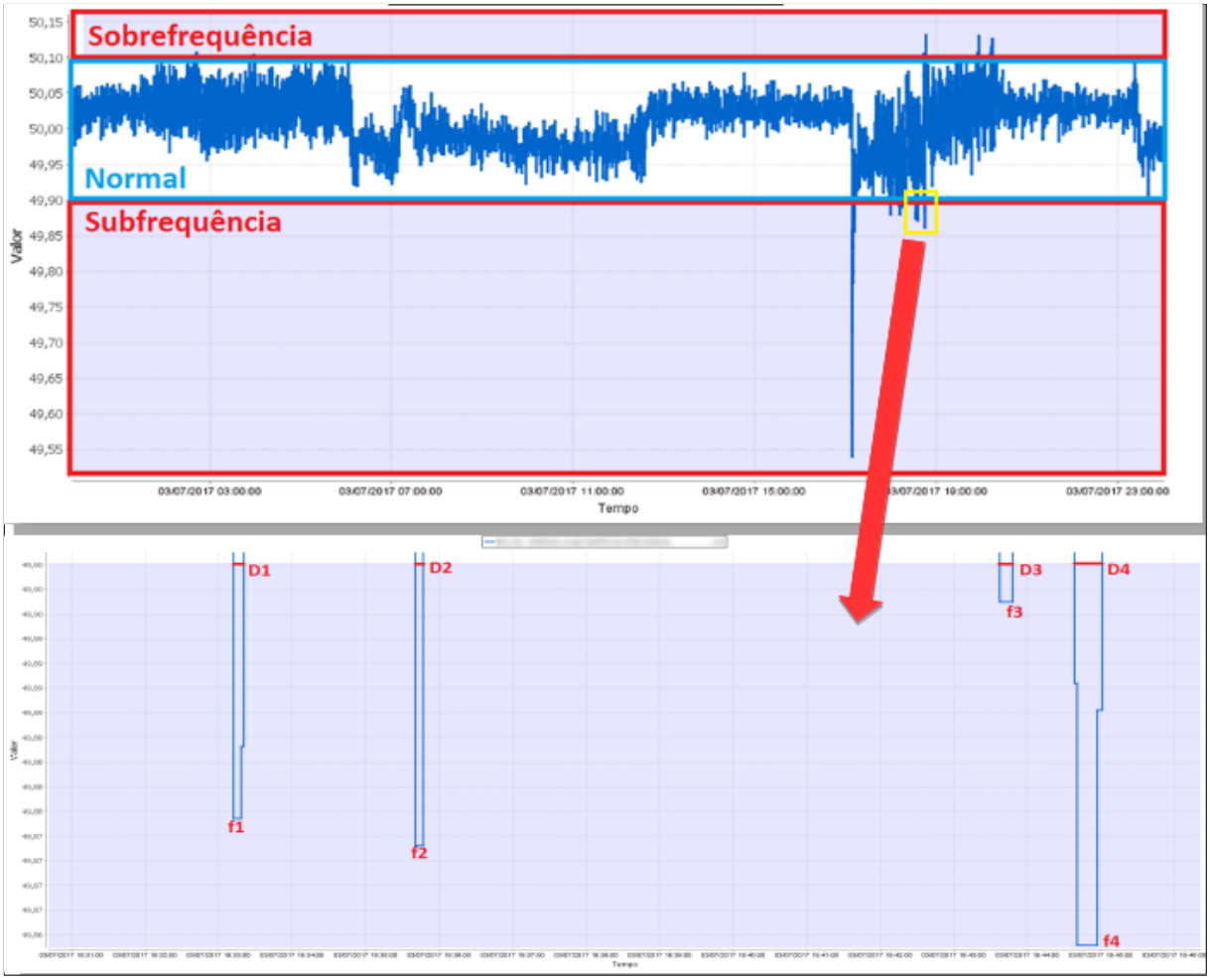

Figura 4.5: Representação gráfica de como é feito o cálculo da frequência e duração.

resultante de uma manutenção, por exemplo.

Por fim, para expressar tendências ascendentes ou descendentes da grandeza analisada, será também calculado e disponibilizado como atributo o coeficiente b da equação da reta $y = a + bx$ que pode ser obtido utilizando-se regressão linear, através da equação:

$$b = \frac{\sum x_i y_i - n\overline{xy}}{\sum {y_i}^2 - n\overline{x}^2} \tag{4.1}$$

Onde x_i e y_i são o valor e a estampa da tempo da amostra i e n é a quantidade de amostras do intervalo.

Em resumo, a Tabela 4.1 mostra uma lista dos atributos atualmente disponíveis na ferramenta computacional desenvolvida nesta Dissertação.

Tabela 4.1: Atributos disponíveis na ferramenta de mineração de dados operativos

Atributo	Forma de Agregação
Número de Amostras	Contagem de Amostras do intervalo
Média	$\frac{1}{n}\sum_{i=1}^{n} a_i$
Moda	Valor mais frequente do intervalo
Desvio Padrão	$\sqrt{\frac{\sum_{i=1}^{N}(x_i-\overline{x})^2}{N-1}}$
Mínimo	Valor mínimo do intervalo
Máximo	Valor máximo do intervalo
Frequência	Contagem do número de transições entre a moda e outros valores
Duração	Tempo decorrido com valores diferentes da moda
Inclinação	$\frac{n\sum X_iY_i-\sum X_i\sum Y_i}{n\sum {X_i}^2-(\sum X_i)^2}$

4.4 Construção do Conjunto de Treinamento Supervisionado

Com um banco de dados (*Data Warehouse*) adequado para realizar o treinamento de algoritmos de classificação, cabe agora a etapa de seleção dos atributos, definição dos parâmetros e execução de classificadores não supervisionados. O objetivo desta etapa é criar um conjunto de dados rotulados, de modo que seja possível executar algoritmos de classificação supervisionados na última etapa do processo.

4.4.1 Seleção de Atributos

É tarefa do especialista selecionar os atributos que serão usados como entrada para os algoritmos de mineração de dados. Para isso, ele faz uso de sua experiência, na operação dos equipamentos e dispositivos do sistema elétrico, e da análise dos dados transformados na etapa anterior, representados através de gráficos e tabelas.

A seleção dos atributos ocorre, na ferramenta computacional, através da marcação dos valores desejados, extraídos na etapa anterior, conforme indica a Figura 4.6. O usuário pode ainda selecionar um ou mais conjuntos de atributos para serem usados como entrada para o algoritmo, como mostra a Figura 4.7. Desta forma, é possível correlacionar, através de modelos gerados pelos algoritmos de mineração de dados, não só atributos de uma mesma grandeza, mas diversos atributos de diversas grandezas. Para isso, o usuário precisa configurar duas extrações distintas, na etapa de Seleção e Transformação dos Dados, e então selecionar para análise estas duas extrações e seus respectivos atributos.

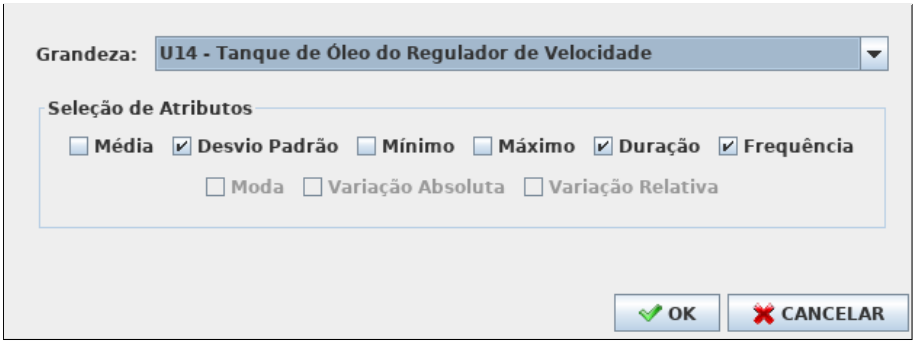

Figura 4.6: Interface gráfica para a seleção de atributos.

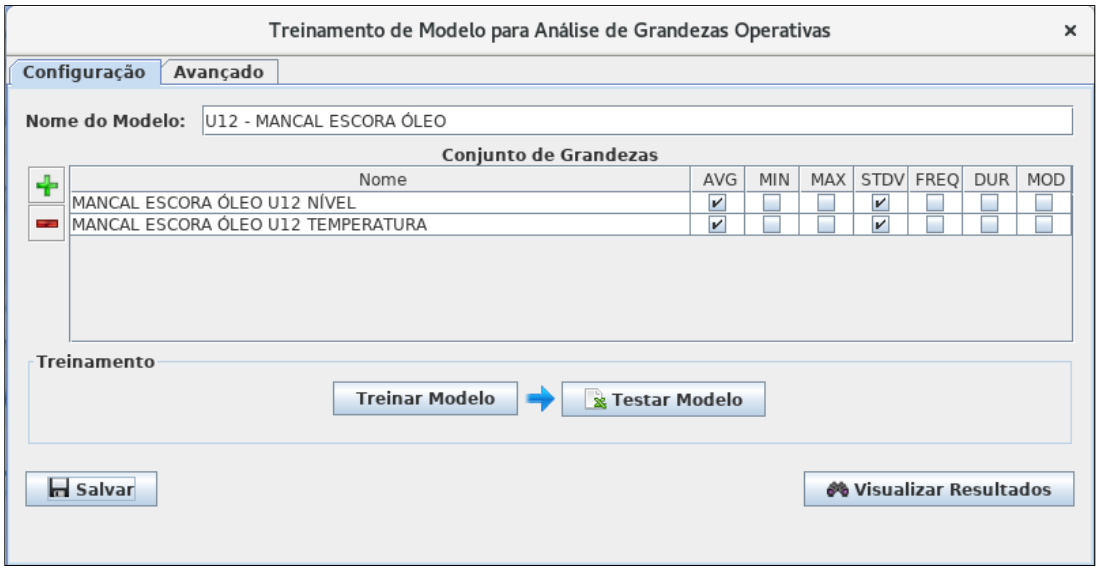

Figura 4.7: Interface gráfica para a configuração e execução do algoritmo de mineração de dados.

4.4.2 Histórico Temporal

Além dos atributos apresentados na seção anterior, calculados utilizando as amostras do intervalo em análise, também serão disponibilizados n valores de instâncias anteriores para compor a lista de atributos. Isto será feito através de um parâmetro, definido pelo usuário, que poderá escolher quantos períodos anteriores ao atual irão compor a lista de atributos de cada instância. Para isso, o sistema percorre a lista de instâncias anteriores em busca destes atributos e os concatena na instância atual. Caso não encontre algum atributo, devido a aplicação de um filtro por exemplo, esta instância (linha) será descartada.

A adição de atributos de estampas de tempo anteriores em uma certa instância permite que o modelo relacione o comportamento dinâmico deste atributo em intervalos subsequentes, permitindo a identificação de variações bruscas, por exemplo, mesmo que

os valores em cada intervalo distinto não sejam considerados anormais.

4.4.3 Execução do Classificador não Supervisionado e Revisão dos Rótulos

Nesta etapa, será feita a pré-rotulagem dos dados. Para esta tarefa, foram testados dois métodos distintos de classificadores não supervisionados, especialmente projetados para a classificação de anomalias, ou seja, conjuntos de dados predominantemente de uma única classe (de dados normais), detalhados nas Seções 3.4.1 e 3.4.3.

Após a execução do algoritmo não supervisionado, as instâncias, agora rotuladas, são salvas em um banco de dados (DW) e a interface da ferramenta computacional permite que o especialista faça uma revisão, como mostra a Figura 4.8, e os aprove para serem então usados no treinamento de algoritmos supervisionados. Para a pesquisa deste trabalho, foram avaliados 4 algoritmos (2 não supervisionados e 2 supervisionados), apresentados na Seção 3.4, mas somente o RNN e o kNN foram efetivamente implementados na ferramenta computacional.

Resultado do Modelo

Salvar

Timestamp	Valor	Fonte	Última Alteração	Anomalia
22/11/2016 04:00:00	0,02	RNN	02/06/2018 18:36:46	☐
22/11/2016 05:00:00	0,02	RNN	02/06/2018 18:36:46	☐
22/11/2016 06:00:00	0,02	RNN	02/06/2018 18:36:46	☐
22/11/2016 07:00:00	0,02	RNN	02/06/2018 18:36:46	☐
22/11/2016 08:00:00	0,02	RNN	02/06/2018 18:36:46	☐
22/11/2016 09:00:00	0,02	RNN	02/06/2018 18:36:46	☐
22/11/2016 10:00:00	0,02	RNN	02/06/2018 18:36:46	☐
22/11/2016 11:00:00	0,02	RNN	02/06/2018 18:36:46	☐
22/11/2016 12:00:00	0,02	RNN	02/06/2018 18:36:46	☐
22/11/2016 13:00:00	0,02	RNN	02/06/2018 18:36:46	☐
22/11/2016 14:00:00	0,04	RNN	02/06/2018 18:36:46	☑
22/11/2016 15:00:00	0,02	RNN	02/06/2018 18:36:46	☐
22/11/2016 16:00:00	0,02	RNN	02/06/2018 18:36:46	☐
22/11/2016 17:00:00	0,02	RNN	02/06/2018 18:36:46	☐
22/11/2016 18:00:00	0,02	RNN	02/06/2018 18:36:46	☐
22/11/2016 19:00:00	0,02	RNN	02/06/2018 18:36:46	☐
22/11/2016 20:00:00	0,02	RNN	02/06/2018 18:36:46	☐
22/11/2016 21:00:00	0,02	SUPERVISOR 1	02/06/2018 18:37:29	☑
22/11/2016 22:00:00	0,02	RNN	02/06/2018 18:36:46	☐
22/11/2016 23:00:00	0,02	RNN	02/06/2018 18:36:46	☐
23/11/2016 00:00:00	0,02	RNN	02/06/2018 18:36:46	☐

Figura 4.8: Tela de revisão do resultado do algoritmo não-supervisionado e rotulagem do especialista.

4.5 Execução Periódica de Modelos

Nesta terceira e última etapa, os dados rotulados na etapa anterior ficam disponíveis para o agendamento de execução periódica de algoritmo de classificação supervisionado, de acordo com a janela de tempo definida. Com isso, novos dados são periodicamente transformados e agregados ao banco de dados, e processados pelos modelos, que podem gerar alertas para equipe de operação e também alimentar outras tabelas do *DW*, de modo a servir como fonte de dados para outras análises. O agendamento da execução periódica será feito através da tela mostrada na Figura 4.9. Nesta tela, o usuário poderá ativar tal execução, que ocorrerá de acordo com os parâmetros definidos na etapa de configuração.

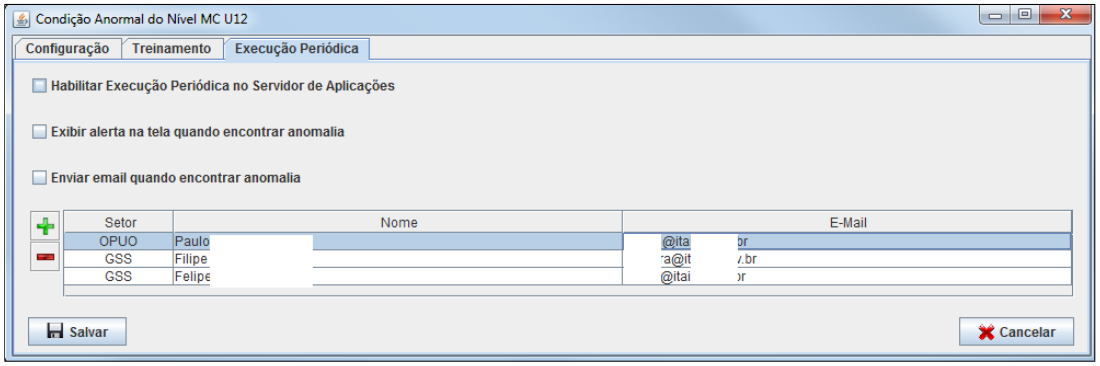

Figura 4.9: Interface para agendamento da execução periódica.

Por fim, o diagrama da Figura 4.10 apresenta uma visão geral da arquitetura da metodologia proposta para a mineração de séries temporais, conforme as etapas descritas anteriormente. Além disso, apresenta, na forma de linhas tracejadas, as interconexões existentes entre as diversas etapas.

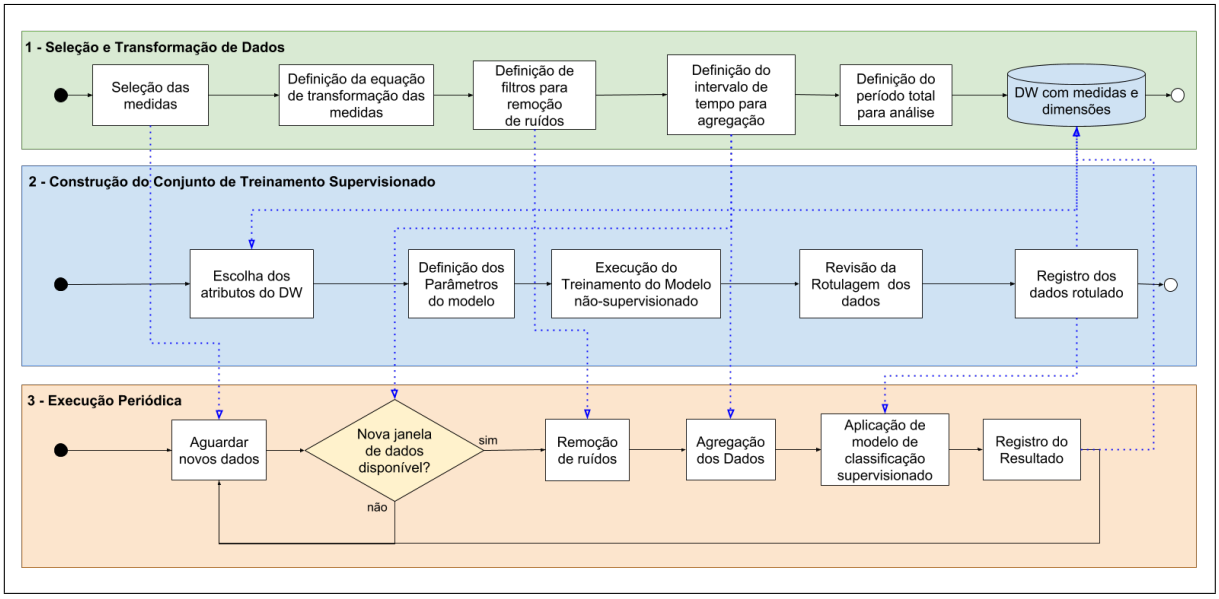

Figura 4.10: Diagrama de atividades da metodologia proposta.

4.6 Implementação da Metodologia Proposta

Como já foi mencionado em diversos exemplos deste Capítulo, uma ferramenta computacional para mineração de dados operativos foi construída, com base na metodologia proposta, para permitir avaliar o seu real potencial. Esta ferramenta, chamada de Assistente para Supervisão de Grandezas Operativas - ASGO, foi desenvolvida em JAVA, utilizando um framework de redes neurais, para implementar as redes replicantes, chamado Neuroph [22]. Os dados pré-processados e os rótulos aplicados pelos algoritmos não supervisionados ou pelos especialistas foram armazenados em um SGBD PostgreSQL. Criou-se também uma biblioteca padronizada de acesso a séries temporais, para acessar não só dados históricos do sistema SCADA como também dados de outros sistemas de apoio, como o sistema de solicitação de serviços e de intervenções da manutenção.

A Figura 4.11 apresenta a tela principal da ferramenta, contendo botões para as três etapas principais apresentadas. Além disso, uma ferramenta para a visualização gráfica das séries temporais também foi desenvolvida para auxiliar na etapa de seleção das medidas que serão usadas. A visualização dos atributos extraídos e seus rótulos é feita utilizando uma ferramenta de análise de dados que se conecta ao *BD*, chamada *Tableau*.

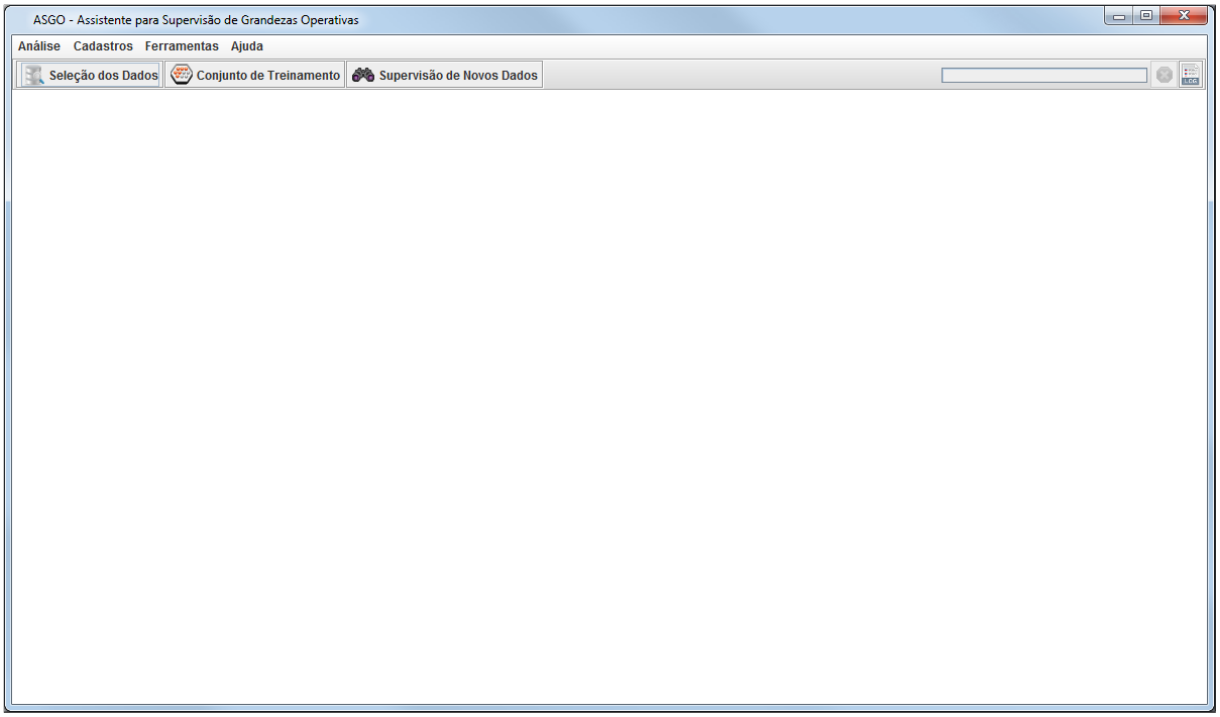

Figura 4.11: Tela principal da ferramenta de mineração de dados operativos.

Capítulo 5

Resultados

5.1 Introdução

Neste capítulo serão apresentados os resultados de alguns estudos de casos realizados na usina hidrelétrica de Itaipu, relativos à aplicação da metodologia proposta no Capítulo 4, em um ambiente real de produção de energia elétrica. Tais estudos se voltam para a avaliação da ferramenta computacional desenvolvida, tanto em termos gerais (aplicação na operação de sistemas de potência), como também sob aspectos particulares da aplicação em estudo. Além disso, os casos foram elaborados seguindo uma ordem de complexidade: dos mais simples até os complexos. O usuário da ferramenta, especialista do sistema elétrico de potência, pode executar as seguintes tarefas: configurar a seleção e transformação dos dados (extração); realizar a criação do conjunto de treinamento supervisionado; por fim, estabelecer a execução periódica de um classificador supervisionado para novos dados.

5.2 Método de Avaliação dos Resultados

Como forma de avaliação, os resultados dos quatro algoritmos de mineração de dados (supervisionados ou não), descritos na Seção 3.4, serão submetidos à chamada *k-fold cross validation*, que divide o conjunto de dados em k subconjuntos e, iterativamente, usa um destes conjuntos como teste e os outros $k - 1$ conjuntos para o treinamento, até que todos os conjuntos tenham sido testados. O valor de $k = 10$ será adotado [16]. Além da acurácia, métricas como *precisão* e *recall* serão utilizadas para a avaliação dos algoritmos e do seu desempenho em relação a diagnósticos ditos falsos positivos e falsos negativos. Estas duas últimas métricas, definidas pela equação 5.1, são calculadas apenas para o conjunto de anomalias.

$$\text{Precisão} = \frac{tp}{tp + fp}, \ \text{Recall} = \frac{tp}{tp + fn} \tag{5.1}$$

Onde tp são as instâncias classificadas corretamente como positivas (verdadeiros positivos), fp são as instâncias classificadas incorretamente como positivas (falsos positivos) e fn são as instâncias classificadas incorretamente como negativas (falsos negativos).

Observa-se que, devido ao número significativamente menor de instâncias de classe anômala do que normal, é natural que a acurácia seja sempre alta. Por isso, é importante não avaliar os resultados exclusivamente por esta métrica.

Em todos os testes apresentados, os dados utilizados foram extraídos por meio da ferramenta de mineração de dados operativos, desenvolvida consoante à metodologia proposta. Os dados pré-processados e os rótulos aplicados pelos algoritmos não supervisionados e revisados pelos especialistas foram armazenados em um banco de dados. Em seguida, estes dados foram normalizados e exportados para a ferramenta *Weka*, que foi utilizada para avaliar os algoritmos de mineração de dados utilizados neste trabalho, aplicando a validação cruzada e calculando as métricas apresentadas nesta Seção [11].

As implementações existentes no repositório do Weka para o *KNN*, *J48* e *Isolation Forest* foram usadas. Além disto, implementou-se um novo pacote para *Weka* com o classficador baseado em rede neural replicante *RNN*, pois ainda não existia implementação deste modelo. Este novo pacote para *Weka* foi desenvolvido com o mesmo código usado na ferramenta de mineração de dados operativos, também em linguagem Java.

5.3 Usina de Itaipu

A Usina Hidrelétrica de Itaipu localiza-se no Rio Paraná, na fronteira entre o Brasil e o Paraguai, região dos municípios de Foz do Iguaçu e Ciudad Del Este, como indica a Figura 5.1. Tal usina foi constituída a partir da assinatura do Tratado de Itaipu, firmado entre estes dois países em 1973. Entre suas características notáveis estão: potência instalada de 14.000MW; 20 unidades geradoras que individualmente são capazes de produzir 700MW; liderança mundial na produção de energia elétrica, alcançando 103.068.366MWh em 2016 e aproximadamente 2,5 bilhões de MWh, desde o início de sua operação em maio de 1984. A Figura 5.2 ilustra a produção anual da Usina de Itaipu. Devido ao porte, esta Usina participa significativamente na produção brasileira e paraguaia de energia elétrica, exercendo papel importante na manutenção da estabilidade dos sistemas interligados destes

dois países.

Figura 5.1: Localização geográfica da Usina de Itaipu.

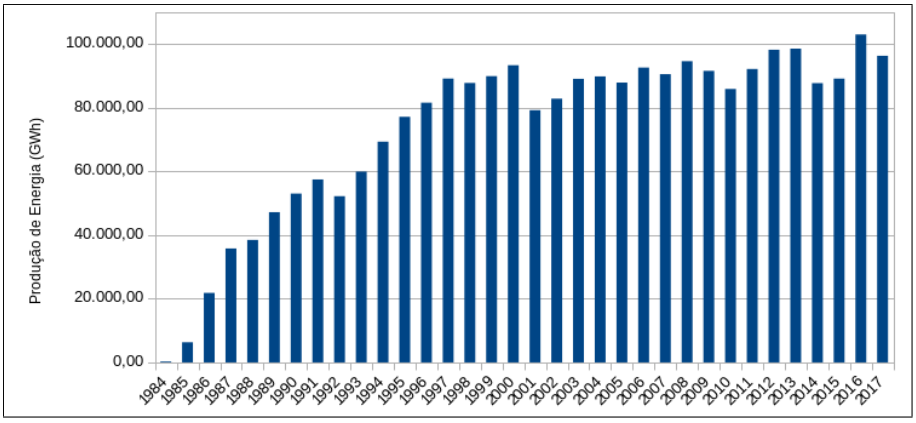

Figura 5.2: Evolução anual da produção de energia da Usina de Itaipu desde sua entrada em operação.

De acordo com o tratado binacional, divide-se a energia produzida em Itaipu igualmente entre Brasil e Paraguai, sendo transmitida aos dois países através de linhas de transmissão de 500kV, conectadas aos respectivos sistemas interligados. Além disso, o tratado também prevê que, caso um dos dois países não consuma toda a sua parte da energia, o excedente será comercializado para o país parceiro. Porém, como a frequência elétrica do sistema elétrico brasileiro é de 60Hz e a do sistema paraguaio de 50Hz, a usina de Itaipu possui dois setores distintos, a saber: as dez primeiras unidades, chamadas de U01 até U9A, localizadas na margem direita da barragem, produzem energia em 50Hz;

e as outras dez unidades, chamadas de U10 até U18A, produzem energia em 60Hz. Por isso, para que o Brasil seja capaz de receber a energia excedente do Paraguai (produzida em 50Hz), esta é retificada para corrente contínua, em uma subestação de Furnas em Foz do Iguaçu - PR, e transmitida através de um elo de corrente contínua de 600kV até Ibiúna-SP, onde então converte-se para 60Hz. A energia produzida no setor de 60Hz, com a tensão elevada para 765kV nesta subestação de Furnas, segue para Ivaiporã, Itaberá e Tijuco Preto, como mostra a Figura 5.3.

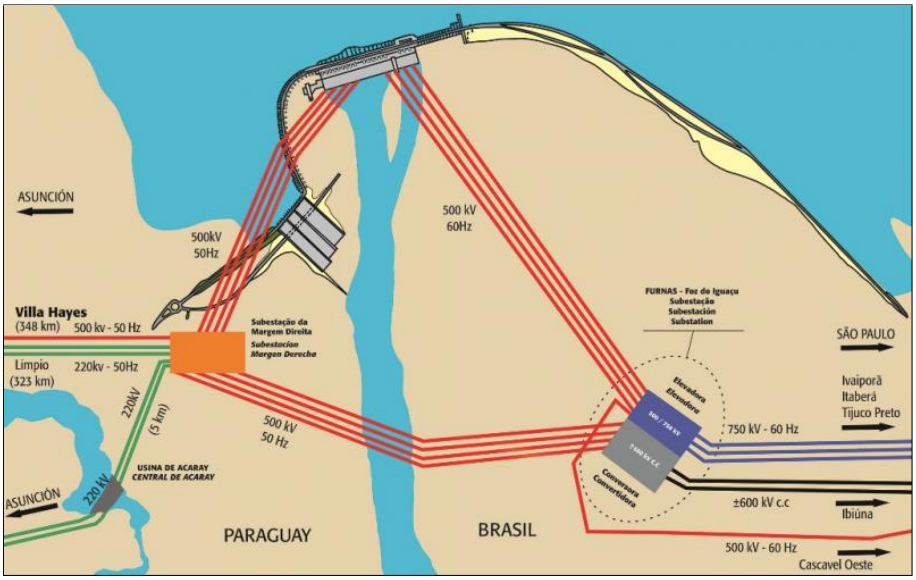

Figura 5.3: Sistema de transmissão da energia elétrica produzida em Itaipu.

5.3.1 A Operação de Itaipu

A operação da usina é realizada pela Superintendência de Operação, que tem um modelo de gestão baseado em três pilares fundamentais: a segurança das pessoas e meio ambiente, a produção de energia e a segurança dos equipamentos e patrimônio. É formada por dois departamentos com responsabilidades distintas. O Departamento de Operação de Usina (OPU), que é responsável por toda a operação dos equipamentos da usina, e o Departamento de Operação do Sistema (OPS), responsável pela operação do sistema, regulando o intercâmbio de energia entre a usina e os agentes responsáveis (Eletrobras no Brasil e Ande no Paraguai). Entre as atividades do Departamento de Operação de Usina, está a supervisão e controle das unidades geradoras que atualmente é feito com o apoio de sistemas digitais.

As próximas Seções apresentam alguns casos reais, propostos por especialistas da OPU, e casos simulados para testar características da ferramenta, inspirados nestes casos

reais.

5.4 Estudo de Caso 1: Simulação de Ponto de Estado

Este primeiro estudo de caso apresenta uma simulação de um ponto de estado que representa o funcionamento de uma bomba de água para remoção de infiltração, acionada a cada 1 minuto, alternando seu valor de estado durante 31 dias, totalizando 44.641 amostras de dados brutos. Em alguns instantes, dentro de todo o período, foram simuladas anomalias em que o valor atual do estado do equipamento permanece por um tempo significativamente maior que o normal. Este comportamento é característico de uma bomba que remove água de infiltração, por exemplo. Quando a infiltração é maior do que o normal, a bomba fica mais tempo ligada para remover o volume de água maior.

A Figura 5.4 mostra um exemplo de anomalia simulada, que ocorreu a partir das 15:00 e durou até as 16:30, durante o funcionamento normal da bomba. Como já foi comentado, no sistema SCADA, os estados são representados por valores numéricos e, no caso desta simulação, 0 indica desligado e 5 indica ligado[1].

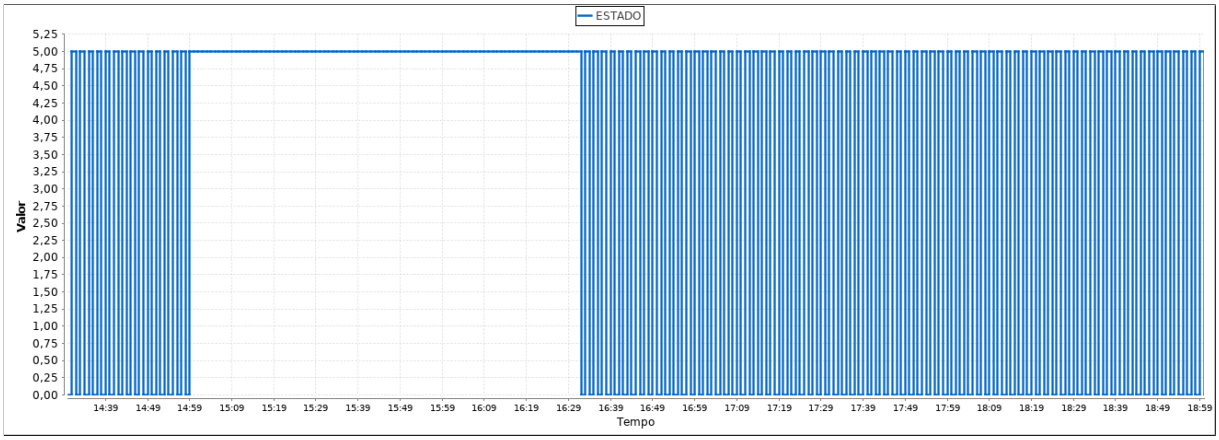

Figura 5.4: Anomalia simulada (permanecendo entre 15:00 h e 16:30 h) durante o funcionamento normal da bomba.

Os dados foram extraídos através da opção de uma análise de Estado e considerando o valor do estado da própria série de entrada (bomba) como equação de transformação, sem nenhum filtro aplicado. A frequência e duração das transições de estado foram então calculadas considerando eventos com valores diferentes da moda. Um intervalo de agregação de 1 hora foi selecionado, gerando uma base de dados com 744 registros, contendo 7

[1]Os valores 0 e 5 foram escolhidos aleatoriamente e quaisquer outros dois valores distintos poderiam ser usados.

anomalias distintas (aproximadamente 1%). A Tabela 5.1 resume a configuração utilizada neste estudo de caso.

Tabela 5.1: Parâmetros do Estudo de Caso 1.

Descrição	Valor
Tipo de Análise	Estado
Equação principal	ESTADO DA BOMBA
Filtro	Nenhum
Intervalo de Agregação	1 Hora
Período de Análise	01/01/2018 até 31/01/2018
Atributos Selecionados	Frequência e Duração
Instâncias	744

Para a construção do conjunto de treinamento supervisionado, foram selecionados, após análise gráfica de seus valores, os atributos frequência e duração em cada estado. Em seguida, foi aplicado o algoritmo não supervisionado, para realizar a pré-rotulagem dos dados e, posteriormente, realizou-se uma revisão para identificar corretamente os períodos anômalos que, neste primeiro caso simulado, não identificou nenhuma instância classificada incorretamente. Por fim, os dados, agora rotulados mostrados na Figura 5.5, foram exportados para um formato compatível com o *Weka* para execução e avaliação dos 4 algoritmos apresentados.

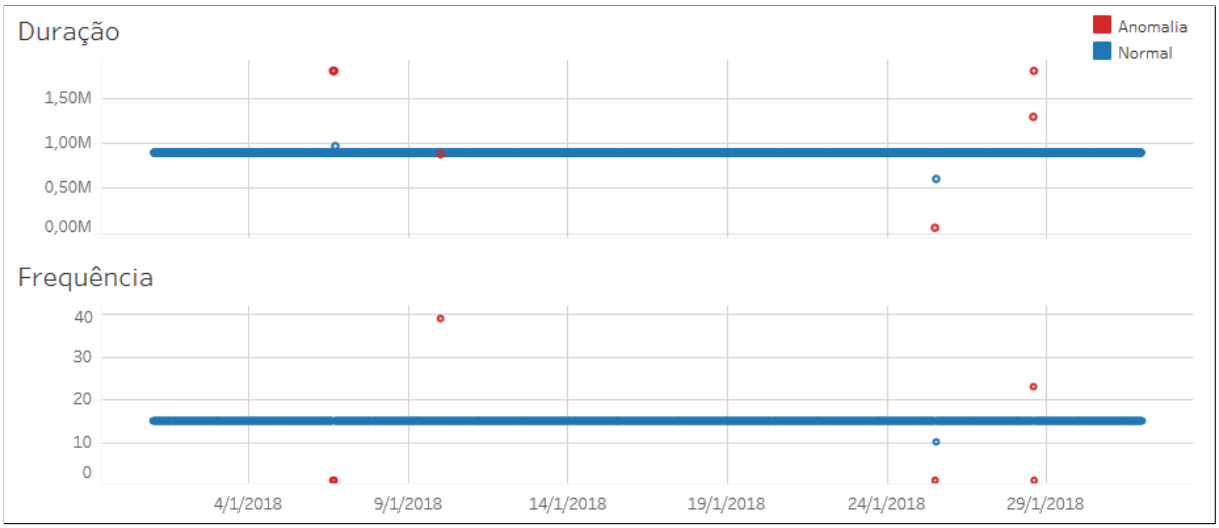

Figura 5.5: Atributos de frequência e duração calculados após transformação de dados.

Os resultados da Tabela 5.2 mostram que os classificadores não supervisionados apresentaram uma boa resposta. O RNN classificou corretamente todas as instâncias, o que já era esperado, considerando que não foi preciso fazer nenhuma alteração no conjunto de treinamento após a etapa de pré-rotulagem (revisão do especialista). Entre os al-

Tabela 5.2: Resultado do Estudo de Caso 1.

Algoritmo	Parâmetros	Acurácia	Precisão	Recall
RNN	D=0,3	100%	100%	100%
Isolation Forest		99,9%	85,7%	85,7%
KNN	k=1	99,9%	100%	71,4%
J48	SMOTE	99,9%	75,0%	85,7%

goritmos supervisionados, o KNN também teve bom desempenho. O J48 apresentou, inicialmente, um resultado bastante desfavorável. Por isso, foi realizado um balanceamento entre classes, utilizando a ferramenta *Weka* para gerar novas instâncias de classes anômalas, baseadas nas existentes, e o resultado foi muito melhor.

Observa-se ainda que os resultados apresentados na Tabela 5.2 são os melhores encontrados para cada algoritmo. Além do balanceamento entre classes (SMOTE) para o J48, descrito na Seção 3.3.3, foi feita a variação dos parâmetros de alguns algoritmos, como o número de vizinhos (k) para o *KNN* e o limite para a distância entre a saída prevista e a entrada (D) para o *RNN*. Esta mesma abordagem foi usada para todos os estudos de casos subsequentes.

5.5 Estudo de Caso 2: Bomba de Óleo do Sistema Regulador de Velocidade

Este estudo de caso representa um caso real de anomalia (ver Figura 5.6) que pode ser detectada através da supervisão de um ponto de estado, como o apresentado na simulação da Seção 5.4. Durante as atividades de inspeção, realizada pelos operadores em uma unidade geradora, pode-se citar a avaliação da atuação histórica da bomba de reposição do óleo do sistema regulador de velocidade das unidades geradoras.

Durante o funcionamento do sistema de regulação de velocidade das unidades geradoras, é normal que ocorra a perda de uma pequena quantidade de óleo, drenado para o tanque de infiltração do regulador de velocidade. Então, quando o nível de óleo do sistema sofre redução significativa, uma bomba que atua neste tanque é acionada para a reposição de óleo. Ocorre que em situações anômalas, caracterizadas pelo vazamento excessivo de óleo do sistema, esta bomba passa a ser ativada em uma frequência maior do que a usual. Desta forma, através de inspeções periódicas no funcionamento da bomba, os operadores conseguem identificar possíveis vazamentos e evitar danos maiores ao subsistema da unidade geradora.

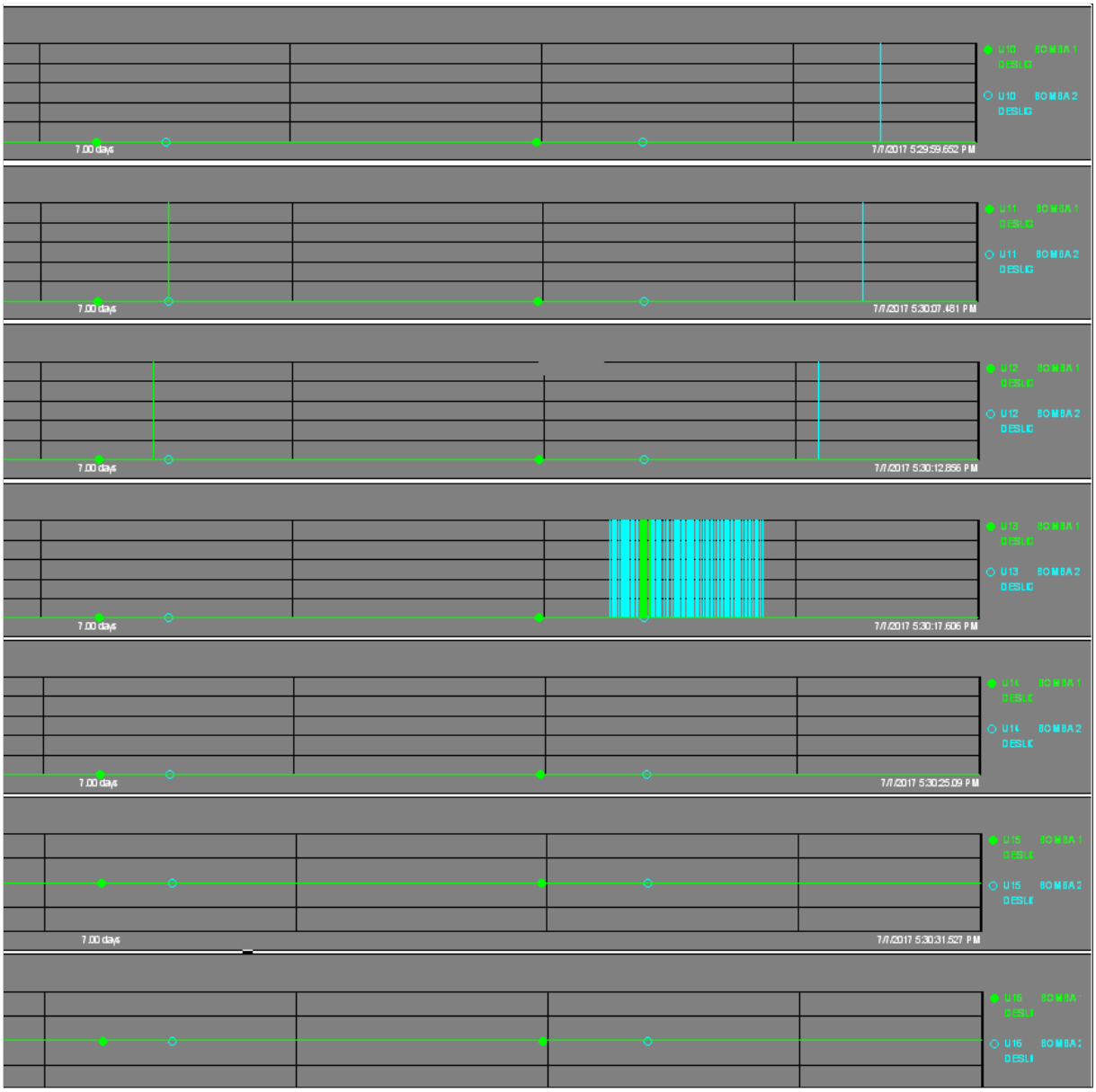

Figura 5.6: Exemplo de anormalidade na quantidade de partidas da bomba do regulador de velocidade da U13.

Na Figura 5.6, pode-se observar que a bomba da U13, localizada no quarto gráfico de cima para baixo, atua (muda de estado) em uma frequência significativamente maior do que as bombas das demais unidades. Portanto, apesar da ativação desta bomba e da perda de óleo do sistema serem fenômenos normais do processo, sua dinâmica pode indicar possíveis anomalias que, atualmente, são identificadas por inspeções manuais nos dados operativos disponíveis.

Para este estudo de caso real, foram obtidos 24 meses de dados de estado (ativo, inativo) das duas bombas de reposição do óleo do regulador de velocidade. Em seguida, os dados foram transformados (agregados) em função da frequência de ativação diária

da bomba; foram coletadas 729 instâncias, contendo 16 anomalias (aproximadamente 2%). Como existem duas bombas operando em paralelo neste sistema, foram configuradas duas extrações de dados distintas, que foram posteriormente relacionadas na etapa de construção do conjunto de treinamento supervisionado, através da estampa de tempo de seus registros. Cada uma das extrações foi configurada conforme a Tabela 5.3.

Tabela 5.3: Parâmetros do Estudo de Caso 2.

Descrição	Valor
Tipo de Análise	Estado
Equação principal	Estado da Bomba
Filtro	Nenhum
Intervalo de Agregação	24 Horas
Período de Análise	01/01/2016 até 31/12/2017
Atributos Selecionados	Frequências das Bombas 1 e 2
Instâncias	727

Foram selecionados, após análise gráfica de seus valores, os atributos frequência das duas bombas (extrações) para a construção do conjunto de treinamento supervisionado. Posteriormente, foi aplicado o algoritmo não supervisionado para realizar a pré-rotulagem dos dados e, em seguida, realizou-se uma revisão para identificar corretamente os períodos anômalos, alterando a classificação de apenas uma instância. Por fim, os dados rotulados, mostrados na Figura 5.7, foram exportados para um formato compatível com o *Weka* para execução dos algoritmos apresentados.

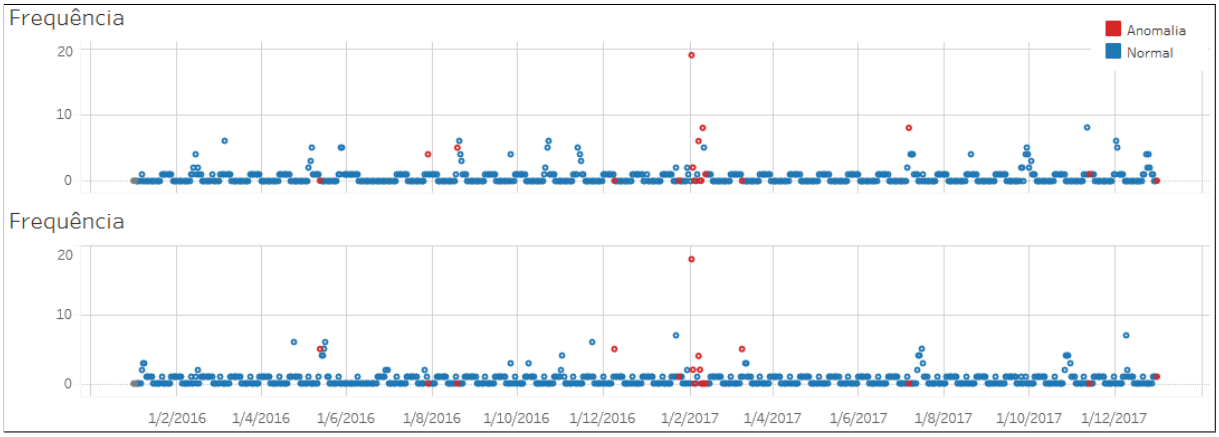

Figura 5.7: Atributos de Frequência das Bombas 1 (acima) e 2 (abaixo).

A Tabela 5.4 apresenta os resultados obtidos com a aplicação dos algoritmos para o estudo de caso 2. O RNN apresentou resultados satisfatórios, mantendo um equilíbrio entre as métricas. O Isolation Forest foi o que obteve o melhor desempenho em identificar instâncias anômalas, encontrando todas as 16 instâncias, porém, apresentou um

Tabela 5.4: Resultado do Estudo de Caso 2.

Algoritmo	Parâmetros	Acurácia	Precisão	Recall
RNN	D=0,11	98,5%	61,9%	81,3%
Isolation Forest		87,5%	15,0%	100%
KNN	k=1	98,5%	85,7%	37,5%
J48	SMOTE	97,8%	50,0%	43,8%

desempenho insatisfatório relacionado à precisão, classificando quantidade significativa de instâncias normais como anomalias. O KNN apresentou resultado satisfatório, porém, deixou de classificar corretamente algumas anomalias. Por fim, o J48 não apresentou resultado satisfatório, mesmo sendo realizado o balanceamento entre as classes antes de executá-lo.

5.6 Estudo de Caso 3: Simulação de Correlação Entre Duas Grandezas Físicas

Esta simulação demonstra a capacidade da metodologia de identificar anomalias relacionadas ao desvio de uma variável que tenha forte correlação com outra. Para isso, foram criadas duas séries simuladas com este comportamento, conforme mostra a Figura 5.8.

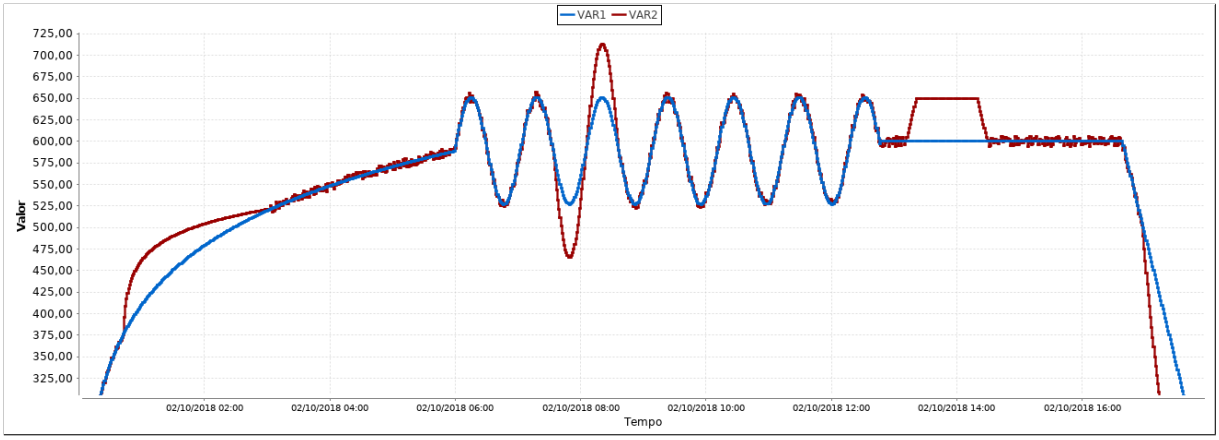

Figura 5.8: Anomalia simulada entre duas grandezas fortemente correlacionadas.

Uma extração foi configurada com a opção de realizar uma análise analógica, registrar o desvio absoluto entre estas duas séries e agregá-lo a cada 15 minutos, durante 4.620 minutos, resultando em 212 registros, contendo 57 anomalias distintas. A Tabela 5.5 resume os parâmetros utilizados na configuração da extração deste estudo de caso.

Tabela 5.5: Parâmetros do Estudo de Caso 3.

Descrição	Valor
Tipo de Análise	Analógica
Equação principal	| A - B |
Filtro	Nenhum
Intervalo de Agregação	15 Minutos
Período de Análise	01/10/2017 00:00 até 03/10/2017 05:00
Atributos Selecionados	Média, Máximo e Mínimo
Instâncias	212

Após análise gráfica de seus valores, os atributos média, mínimo e máximo, foram selecionados para a construção do conjunto de treinamento supervisionado. Em seguida, para realizar a pré-rotulagem dos dados, foi aplicado o algoritmo não supervisionado e, posteriormente, realizou-se uma revisão para identificar corretamente os períodos anômalos, modificando a classificação de 28 instâncias. Por fim, os dados, agora rotulados mostrados na Figura 5.9, foram exportados para execução e avaliação dos 4 algoritmos no *Weka*.

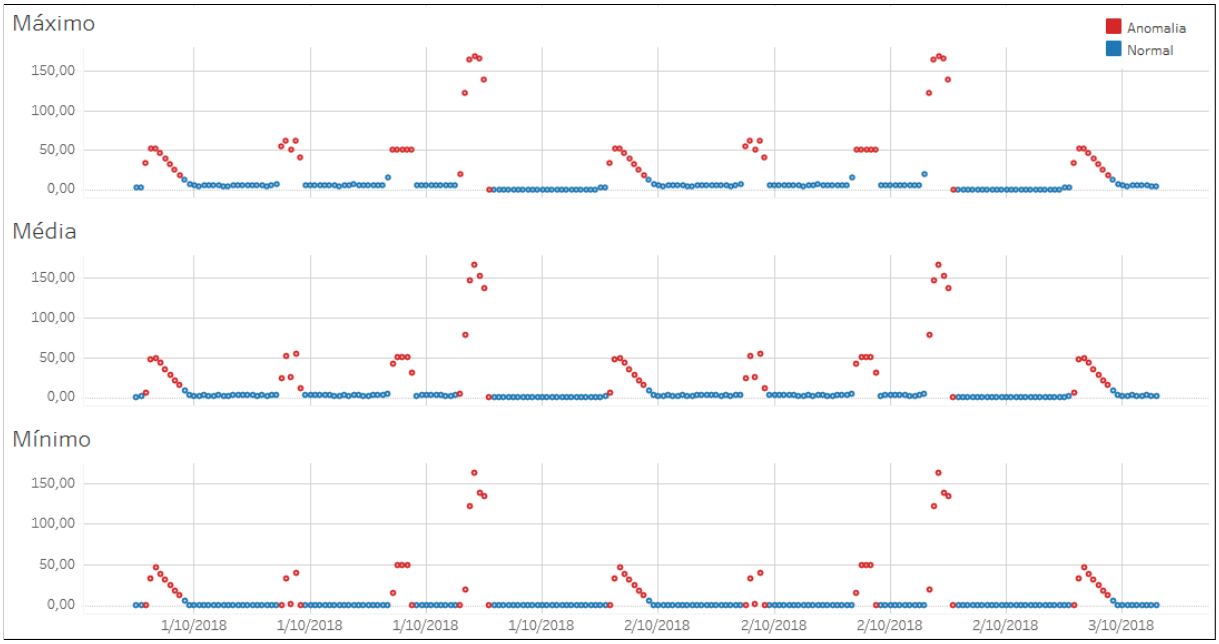

Figura 5.9: Atributos do estudo de caso 3.

A Tabela 5.6 apresenta os resultados do estudo de caso 3, onde observa-se que todos os algoritmos tiveram bons resultados, com o RNN, KNN e o J48 apresentando melhor precisão e o Isolation Forest e o J48 com um maior Recall.

Tabela 5.6: Resultado do Estudo de Caso 3.

Algoritmo	Parâmetros	Acurácia	Precisão	Recall
RNN	D=0,05	96,2%	98,0%	87,7%
Isolation Forest		96,2%	90,2%	96,5%
KNN	k=1	98,1%	98,2%	94,7%
J48		98,5%	98,2%	96,5%

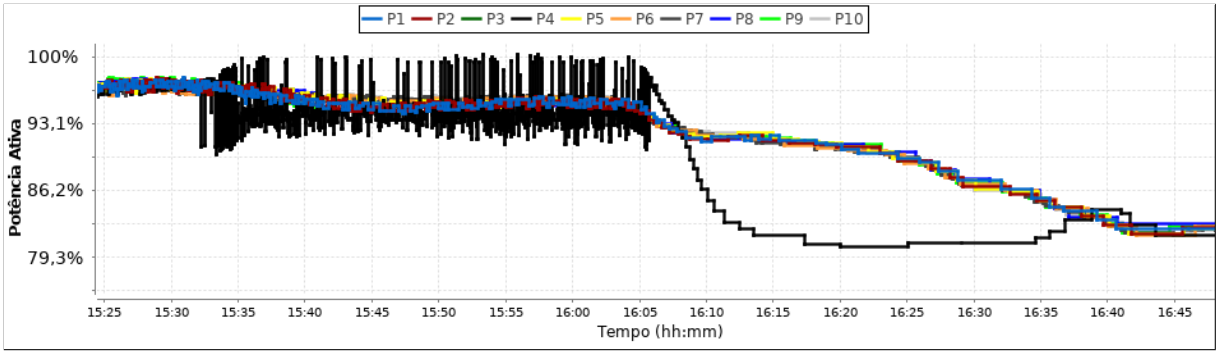

Figura 5.10: Desvio de Potência P4 em relação à potência das outras UGs durante anomalia.

5.7 Estudo de Caso 4: desvio de potência ativa de unidade geradora

Um segundo estudo de caso real, derivado da simulação apresentada na Seção 5.6, corresponde ao desvio de potência ativa de uma unidade geradora em relação à potência ativa das outras unidades, quando elas operam sincronizadas ao Sistema Interligado e em modo de controle automático de geração (CAG). Em algumas situações, são observadas pequenas oscilações de uma unidade em relação às outras, como mostra a Figura 5.10, o que não chega a ativar um alarme, mas representa um comportamento indesejado do sistema.

Para este estudo de caso, foram definidos os parâmetros estabelecidos na Tabela 5.7 e extraídos 3.370 horas de dados, totalizando 5.305 registros, contendo 7 registros rotulados como anômalos durante o período.

Para a construção do conjunto de treinamento supervisionado, foram selecionados os atributos *Máximo* e *Desvio Padrão*. Em seguida, foi aplicado o algoritmo não supervisionado, para realizar a pré-rotulagem dos dados e, então, realizou-se uma revisão para identificar corretamente todos os períodos anômalos, alterando a rotulagem de 4 instâncias. Por fim, os dados rotulados, mostrados na Figura 5.11, foram exportados para um formato compatível com o *Weka* para execução e avaliação dos algoritmos.

Tabela 5.7: Parâmetros do Estudo de Caso 4.

Descrição	Valor
Tipo de Análise	Analógica
Equação principal	$\lvert P_4 - \frac{\sum_{i=1}^{n} P_i}{n} \rvert ; n \neq 4$
Filtro	$P_4 > 0$ & $n > 0$
Intervalo de Agregação	15 minutos
Período de Análise	01/06/2018 até 14/10/2018 23:15
Atributos Selecionados	Máximo e Desvio Padrão
Instâncias	5.305

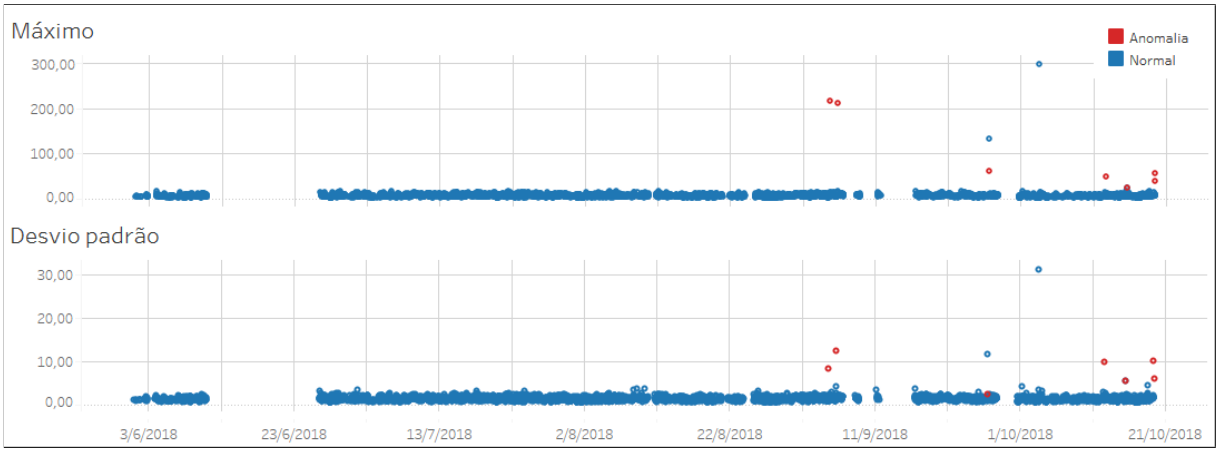

Figura 5.11: Atributos do estudo de caso 4.

Os resultados da aplicação dos algoritmos no *Weka* são apresentados na Tabela 5.8. Observa-se que três dos algoritmos (RNN, KNN e J48) tiveram um bom *recall* e a precisão ficou em torno de 70%. Isto aconteceu porque todos eles classificaram como anômalas variações provocadas por partida e parada repentina de outras unidades, que por alguns segundos tiveram um desvio maior do que o normal antes de mudarem seus estados de operação. Com relação ao Isolation Forest, apesar de ter classificado corretamente todas as anomalias, classificou muitas instâncias normais como anômalas, prejudicando muito sua precisão.

Tabela 5.8: Resultados do estudo de caso 4.

Algoritmo	Parâmetros	Acurácia	Precisão	Recall
RNN	D=0,15	99,9%	75,0%	85,7%
Isolation Forest		82,1%	0,7%	100%
KNN	k=1	99,9%	66,7%	85,7%
J48	SMOTE	99,9%	63,6%	100,0%

5.8 Estudo de Caso 5: Pressão do Tanque de óleo do Sistema Regulador de Velocidade

O quinto estudo de caso trata de um fenômeno real que pode ser observado através da análise do comportamento de um ponto analógico. Além da frequência de ativação da bomba do tanque de óleo do sistema regulador de velocidade, os operadores da usina também acompanham o comportamento da curva de pressão do tanque, que também pode indicar anomalias, conforme mostra a Figura 5.12. Como foi mencionado na Seção 2.2, este óleo em pressão é utilizado para movimentar as palhetas que regulam a abertura do distribuidor, regulando a passagem de água na turbina. Assim, é possível manter a velocidade constante do conjunto turbina-gerador durante variações de carga.

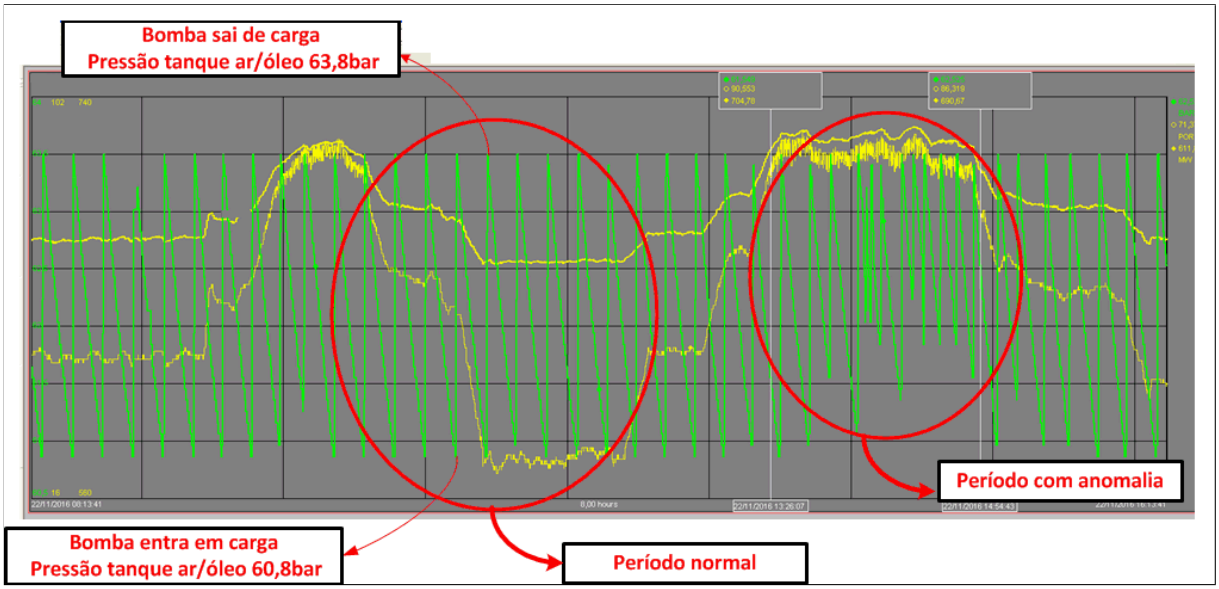

Figura 5.12: Variação anômala da pressão no tanque de óleo do sistema regulador de velocidade da U14 durante o dia 22/11/2016.

Normalmente, variações no comportamento de pressão indicam vazamentos no sistema. Na Figura 5.12 é possível observar que a variação da pressão sofreu significativa perturbação após as 13 horas do dia 22 de novembro de 2016, formando um espaço não preenchido na região inferior do gráfico.

Para a aplicação do método descrito neste trabalho, foi utilizada uma janela de cálculo de uma hora para a agregação dos dados e feita uma análise usando dados de 1 ano, totalizando 7.582 instâncias, contendo 28 instâncias rotuladas como anômalas. A Tabela 5.9 mostra a configuração feita da ferramenta para extração dos dados e construção do conjunto de treinamento rotulado. Observa-se que a variável $PressaoNormaldoSistema$, usada na equação de filtro, é proveniente do historiador mas não identifica as anomalias

buscadas neste estudo de caso, que são contextuais, sendo usada apenas como um filtro para valores muito diferentes dos normais.

Tabela 5.9: Parâmetros do Estudo de Caso 5.

Descrição	Valor
Tipo de Análise	Analógica
Equação principal	$ValorPressao$
Filtro	$PressaoNormaldoSistema \ \& \ Estado = Sincronizado$
Intervalo de Agregação	01 hora
Período de Análise	01/07/2016 00:00 até 01/07/2017 00:00
Atributos Selecionados	Desvio Padrão e Frequência
Instâncias	7.582

Neste estudo de caso, foram selecionados, após análise gráfica de seus valores, os atributos desvio padrão e frequência para a construção do conjunto de treinamento supervisionado. Posteriormente, foi aplicado o algoritmo não supervisionado, para realizar a pré-rotulagem dos dados e, em seguida, realizou-se uma revisão para identificar corretamente todos os períodos anômalos, alterando a classificação de 11 instâncias. Por fim, os dados, agora rotulados, mostrados na Figura 5.13, foram exportados para para execução e avaliação dos 4 algoritmos apresentados.

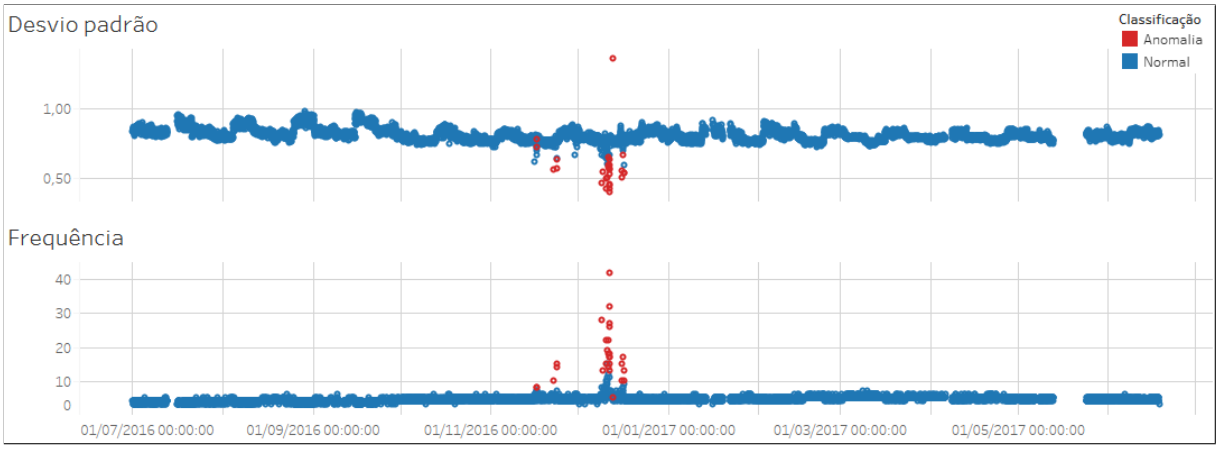

Figura 5.13: Atributos do estudo de caso 5.

Como é possível observar na Tabela 5.10, todos os algoritmos apresentaram boa acurácia e recall. Com relação à precisão, com exceção do Isolation Forest, que mais uma vez classificou muitas instâncias normais como anômalas, todos os outros algoritmos tiveram um bom resultado.

Tabela 5.10: Resultados do estudo de caso 5.

Algoritmo	Parâmetros	Acurácia	Precisão	Recall
RNN	D=0,2	99.9%	92,6%	89,3%
Isolation Forest		81,1%	1,9%	100%
KNN	k=1	99,9%	92,3%	85,7%
J48		99.9%	92,3%	85.7%

5.9 Estudo de Caso 6: Nível de Óleo do Mancal Combinado da Unidade Geradora

O sexto estudo de caso trata de uma anomalia que pode ser observada através da supervisão de dois pontos analógicos. Neste estudo de caso, existe uma correlação entre a temperatura e o nível do óleo do mancal combinado (mancal guia inferior e o mancal de escora), identificado na Figura 5.14, que serve de apoio fixo para os diversos elementos do conjunto gerador.

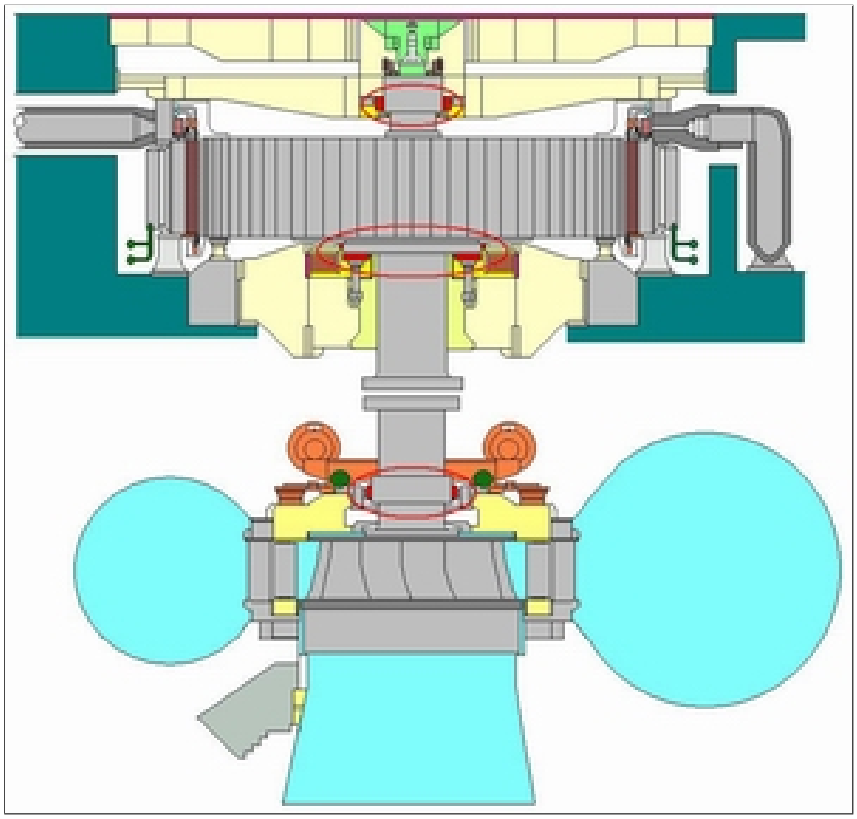

Figura 5.14: Localização dos diversos mancais existentes na Unidade Geradora destacados em vermelho.

A Figura 5.15 apresenta um exemplo da correlação entre as duas grandezas e da anomalia que é objeto deste estudo de caso. A partir de 06/07 é possível observar que

a relação entre as duas grandezas passa a ter um comportamento variável, diferente dos outros períodos em que esta relação é aproximadamente constante quando a UG está em operação.

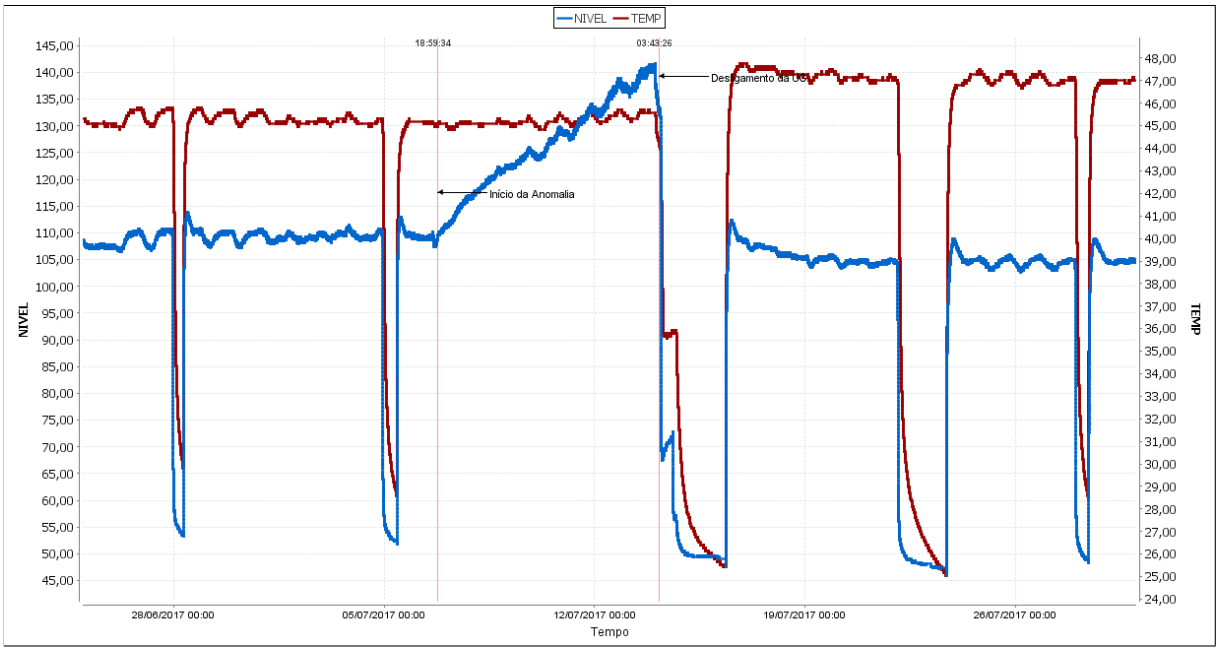

Figura 5.15: Temperatura e nível do óleo do mancal combinado da UG.

Para identificar esta anomalia, foi configurada uma extração com uma janela de cálculo de 24 horas para a agregação e feita uma análise usando dados de 1 ano e meio, totalizando 455 instâncias, após aplicação do filtro da equação e do número mínimo de amostras, contendo 7 instâncias rotuladas como anômalas. A Tabela 5.11 mostra a configuração feita da ferramenta para extração dos dados e construção do conjunto de treinamento rotulado.

Tabela 5.11: Parâmetros do Estudo de Caso 6.

Descrição	Valor
Tipo de Análise	Analógica
Equação principal	$\frac{NIVEL}{TEMPERATURA}$
Filtro	$ESTADO = SINCRONIZADO \ \& \ CONTROLE = AUTO$
Intervalo de Agregação	24 horas
Período de Análise	01/01/2017 00:00 até 01/07/2018 00:00
Atributos Selecionados	Duração e Inclinação
Instâncias	455

Para a construção do conjunto de treinamento supervisionado, foram selecionados os atributos duração e inclinação. Em seguida, foi aplicado o algoritmo não supervisionado,

para realizar a pré-rotulagem dos dados e, então, realizou-se uma revisão para identificar corretamente todos os períodos anômalos, alterando a classificação de 5 instâncias. Por fim, os dados, agora rotulados, mostrados na Figura 5.16, foram exportados para execução e avaliação dos algoritmos apresentados na ferramenta *Weka*.

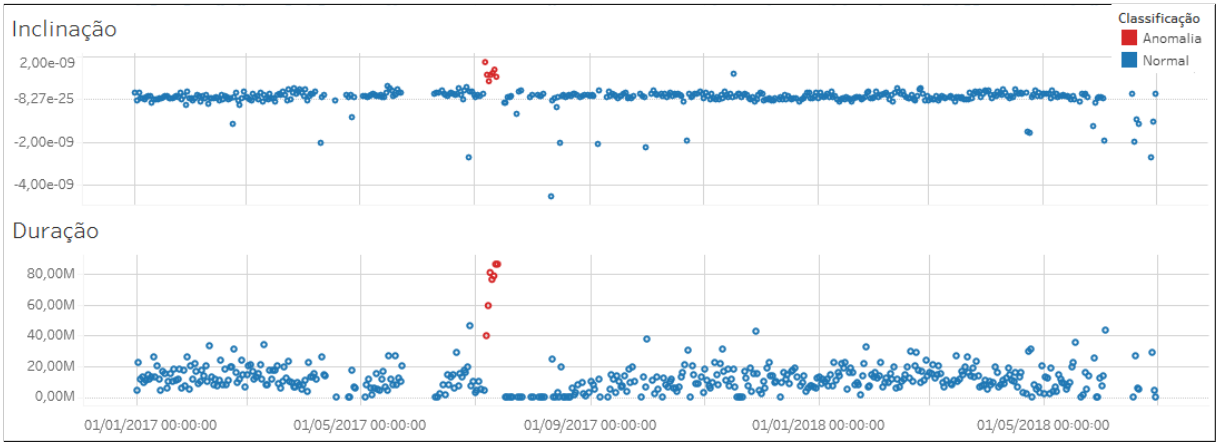

Figura 5.16: Atributos do estudo de caso 6.

Os resultados da aplicação dos algoritmos usando o *Weka* são apresentados na Tabela 5.12. Entre os algoritmos não supervisionados o RNN apresentou um resultado mediano. O Isolation Forest apresentou uma alta precisão mas vários falsos positivos. O KNN e o J48 apresentaram bom resultado, não identificando apenas 1 anomalia, que foi incorretamente classificada como normal.

Considerando o baixo número absoluto de instânicas rotuladas como anomalias neste caso, foi definido um valor de k igual ao número total de instâncias para o método de validação *k-fold*, que é um caso especial da aplicação do método conhecido como *leave-one-out*.

Tabela 5.12: Resultados do estudo de caso 6.

Algoritmo	Parâmetros	Acurácia	Precisão	Recall
RNN	D=0,15	98,9%	62,5%	71,4%
Isolation Forest		90,3%	13,7%	100%
KNN	k=1	100%	100%	85,7%
J48		99,8%	100%	85,7%

5.10 Análise dos Resultados

Nesta Seção, serão apresentadas conclusões relacionadas especificamente aos resultados da aplicação da metodologia nos estudos de casos deste capítulo.

Na etapa de Seleção e Transformação dos dados, observou que os especialistas tiveram rápido entendimento de como proceder para a extração de formas básicas de análise. Foram configuradas com sucesso extrações para representar diversos tipos de casos e feitas extrações para longos períodos de análise. Porém, observa-se que as extrações criadas ainda estão em um estágio inicial de complexidade, pois, com a possibilidade de uso de um número de variáveis ilimitadas, equações e filtros parametrizáveis, extrações futuras podem proporcionar novos tipos de análise ainda não vislumbrados.

Na etapa de construção do conjunto de treinamento supervisionado, o *RNN* apresentou um resultado satisfatório na pré-rotulagem dos dados, sendo, no entanto, necessário ajustar frequentemente o parâmetro de limite de erro. Nos testes realizados, os especialistas não apresentaram dificuldade em ajustar este parâmetro para rotularem a grande maioria dos dados, conforme pretendido.

Já o *Isolation Forest*, apesar de não carecer de ajustes em parâmetros em sua implementação para o *Weka*, que define automaticamente a profundidade máxima para as árvores ($hlim$) com base no número total de instâncias, classificou incorretamente muitas instâncias normais como anomalias. Seu resultado poderia ter sido melhor caso sua implementação disponibilizasse um parâmetro configurável para definir $hlim$.

Com relação à etapa de execução periódica, os dois algoritmos supervisionados apresentaram bons resultados, mas, em alguns casos, precisou ser realizado o balanceamento entre as classes para aplicação do *J48*.

O *KNN*, por sua vez, não necessitou de balanceamento entre classes, e possui ainda a vantagem de permitir uma atualização contínua de sua base de treinamento, através da revisão pelo especialista dos dados classificados. Outra vantagem na aplicação deste tipo de algoritmo, em um fluxo contínuo de novos dados, é que a normalização pode ser feita em tempo de execução, o que facilita o tratamento de novos dados fora dos limites até então conhecidos. O custo computacional para o uso do *KNN* não se mostrou significativo durante os testes realizados.

Assim, no desenvolvimento da ferramenta computacional, optou-se pela implementação do *RNN* para pré-rotulagem dos dados, na etapa de construção do conjunto de treinamento, e do *KNN* para a classificação (supervisionada) de novos dados na etapa de execução periódica.

5.11 Outras Aplicações da Ferramenta

Além do uso para detecção de anomalias, durante a elaboração deste trabalho, surgiram espontaneamente outras aplicações para a ferramenta desenvolvida, relacionadas à extração e transformação dos dados históricos, listadas a seguir:

- Cálculo das propriedades estatísticas do desvio de um novo medidor de nível montante instalado recentemente na usina. Estas informações permitiram identificar outliers e realizar ajustes no medidor durante o seu comissionamento.
- Identificação de falhas no histórico de registro de alarmes, através da contagem do número de amostras médias existentes no banco de dados. Para isso, foi preciso realizar uma verificação em aproximadamente 8 anos de dados que identificou alguns horários sem registros ou com um número muito abaixo do normal.
- Análise de falhas ao registrar (historiar) grandezas analógicas através da comparação do número de amostras e desvio padrão entre duas medidas redundantes (da mesma grandeza física).
- Identificação da ocorrência de um fenômeno de vibração excessiva durante a partida e parada de algumas unidades geradoras.

Assim como os estudos de casos apresentados neste Capítulo, estas atividades extras contribuem para justificar o desenvolvimento deste trabalho, pois trazem benefícios diretos para os usuários que até então não dispunham de uma forma prática de processar todos estes dados.

Capítulo 6

Conclusões

Esta dissertação abordou a aplicação de técnicas de mineração de dados de grandezas operativas de uma usina hidrelétrica, registradas pelo seu sistema de controle e supervisão em um banco de dados histórico. Isto envolve discorrer sobre cada uma das etapas do processo de descoberta de conhecimento e integrá-las de forma coerente, a fim de ter uma metodologia que descreva um processo funcional e prático o suficiente para ser aplicado em um ambiente industrial.

Inicialmente, a necessidade de preparação e transformação dos dados para serem processados pelos algoritmos de mineração de dados se destaca entre as dificuldades encontradas. Ruídos, dados faltantes, diferentes contextos, alterações do contexto e séries com amostras não sincronizadas são exemplos encontrados na maioria dos casos e que precisaram, de alguma forma, ser tratados.

Observou-se que uma solução que integrasse a experiência do especialista e a capacidade de processamento dos sistemas computadorizados traria uma boa relação de custo-benefício, pois concentra no especialista atividades rápidas, porém extremamente complexas, como a seleção de dados, definição da janela de agregação, seleção dos atributos e parametrização básica dos algoritmos de mineração de dados. Por outro lado, delega à máquina a tarefa de processar uma grande quantidade de dados, utilizando estruturas e algoritmos bem estabelecidos. Ou seja, a metodologia proposta permite que os benefícios das técnicas de mineração de dados sirvam como ferramenta de suporte às atividades dos especialistas de uma usina hidrelétrica com uma grande quantidade de dados ainda inexplorados.

Neste sentido, a aplicação de métodos não supervisionados para realizar uma pré-rotulagem dos dados, seguida de uma revisão pelo especialista, mostrou-se como uma

forma promissora de se obter conjuntos de treinamento para os algoritmos supervisionados.

Na etapa de execução periódica, o processamento de um fluxo contínuo de novos dados que, em alguns casos eram diferentes de tudo o que já tinha sido visto, se mostrou um desafio, pois estes novos dados precisaram ser normalizados e aplicados a algoritmos previamente treinados com o conjunto de dados históricos. Neste caso, a flexibilidade proporcionada por algoritmos *lazy* para incorporar no conjunto de treinamento novos dados, inclusive também revisados pelo especialista, também foram resultados desta pesquisa.

De maneira mais ampla, a utilização de uma arquitetura bem estabelecida, com a presença de um *Data Warehouse* e uma estrutura de dados padronizada, proporciona ainda o uso dos resultados de cada etapa do processo por outras formas de análises, como a análise *OLAP* ou outras ferramentas que poderiam ser conectadas a este *Data Warehouse*.

Além da arquitetura, a caracterização dos dados em séries temporais permite a integração de diferentes fontes de dados, como o banco de dados de autorizações de trabalho, solicitações de serviço e alarmes de falhas em equipamentos.

Como contribuição científica deste trabalho, destaca-se a metodologia proposta no Capítulo 4 para a análise de séries temporais de grandezas físicas, analógicas e de estado, agregadas em intervalos de tempo, como forma de transformação para a sua representação em estruturas de dados capazes de serem processadas por algoritmos tradicionais de mineração de dados. Além disso, os resultados das diferentes formas de agregação aplicadas, como desvio padrão, inclinação, frequência e duração em faixas discretas também podem servir como referência para outros trabalhos.

6.1 Trabalhos Futuros

Este trabalho criou a infraestrutura para a realização de outras pesquisas que podem dele ser derivadas. Entre elas, pode-se destacar a avaliação de meta-heurísticas para a seleção automática de grandezas para serem avaliadas, através da utilização de algoritmos genéticos, por exemplo. Para isso, seria necessário definir uma função objetivo capaz de representar a relevância das grandezas selecionadas, o que pode ser uma grande desafio para o caso de grandezas analógicas, mas possivelmente factível para grandezas de estado.

Vislumbra-se também, como possível continuação do trabalho, a investigação de novas formas de agregação dos dados brutos, que evidenciem outras propriedades nos dados

analisados. Além disso, estas novas formas de agregação podem ser facilmente implementadas no aplicativo computacional desenvolvido, por meio de interface que permita a sua inclusão sem necessidade de alterar a sua estrutura. Para tal, pode-se utilizar, inclusive, uma linguagem específica que permita, ao usuário especialista, acrescentar uma nova forma de agregação.

Por fim, existe ainda a possibilidade do desenvolvimento de um sistema especialista capaz de correlacionar múltiplas anomalias detectadas pelo método descrito neste trabalho e apresentar um diagnóstico que identifique as causas do problema encontrado.

Referências

[1] Bezerra, U.; Ohana, I.; Vieira, J. Data-mining experiments on a hydroelectric power plant. *IET Generation, Transmission & Distribution 6*, 5 (2012), 395–403.

[2] Boyer, S. A. *SCADA supervisory control and data acquisition.* The Instrumentation, Systems and Automation Society, 2018.

[3] Chandola, V.; Banerjee, A.; Kumar, V. Anomaly detection: A survey. *ACM computing surveys (CSUR) 41*, 3 (2009), 15.

[4] Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; Kegelmeyer, W. P. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research 16* (2002), 321–357.

[5] Cheboli, D. *Anomaly detection of time series.* Tese de Doutorado, University of Minnesota, 2010.

[6] Cover, T.; Hart, P. Nearest neighbor pattern classification. *IEEE transactions on information theory 13*, 1 (1967), 21–27.

[7] Dau, H. A.; Ciesielski, V.; Song, A. Anomaly detection using replicator neural networks trained on examples of one class. In *Asia-Pacific Conference on Simulated Evolution and Learning* (2014), Springer, pp. 311–322.

[8] Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P. From data mining to knowledge discovery in databases. *AI magazine 17*, 3 (1996), 37.

[9] Goldstein, M.; Uchida, S. Behavior analysis using unsupervised anomaly detection. In *The 10th Joint Workshop on Machine Perception and Robotics (MPR 2014). Online* (2014).

[10] Gupta, M.; Gao, J.; Aggarwal, C. C.; Han, J. Outlier detection for temporal data: A survey. *IEEE Transactions on Knowledge and Data Engineering 26*, 9 (2014), 2250–2267.

[11] Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I. H. The weka data mining software: an update. *ACM SIGKDD explorations newsletter 11*, 1 (2009), 10–18.

[12] Han, J.; Pei, J.; Kamber, M. *Data mining: concepts and techniques.* Elsevier, 2011.

[13] Hawkins, D. M. *Identification of outliers*, vol. 11. Springer, 1980.

[14] Hawkins, S.; He, H.; Williams, G.; Baxter, R. Outlier detection using replicator neural networks. *Proceedings of the 4th International Conference on Data Warehousing and Knowledge Discovery* (2002), 170–180.

[15] King, S.; King, D.; Astley, K.; Tarassenko, L.; Hayton, P.; Utete, S. The use of novelty detection techniques for monitoring high-integrity plant. In *Control Applications, 2002. Proceedings of the 2002 International Conference on* (2002), vol. 1, IEEE, pp. 221–226.

[16] Kohavi, R., et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai* (1995), vol. 14, Montreal, Canada, pp. 1137–1145.

[17] Koychev, I. Gradual forgetting for adaptation to concept drift. In *Current Issues in Spatio-Temporal Reasoning* (2000), Proceedings of ECAI 2000 Workshop,.

[18] Liu, F. T.; Ting, K. M.; Zhou, Z.-H. Isolation forest. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on* (2008), IEEE, pp. 413–422.

[19] Patcha, A.; Park, J.-M. An overview of anomaly detection techniques: Existing solutions and latest technological trends. *Computer networks 51*, 12 (2007), 3448–3470.

[20] Quinlan, J. R. *C4. 5: programs for machine learning*. Elsevier, 2014.

[21] Salvador, S.; Chan, P.; Brodie, J. Learning states and rules for time series anomaly detection. In *FLAIRS Conference* (2004), pp. 306–311.

[22] Sevarac, Z.; Goloskokovic, I.; Tait, J.; Carter-Greaves, L.; Morgan, A.; Steinhauer, V. Neuroph-java neural network framework. *Retrieved in January* (2012).

[23] Widmer, G.; Kubat, M. Learning in the presence of concept drift and hidden contexts. *Machine learning 23*, 1 (1996), 69–101.

[24] WIERINGA, R. J. *Design Science Methodology for Information Systems and Software Engineering*. Springer Berlin Heidelberg, 2014.