Geração Semiautomática de Valores de Referência para Identificação de Obstruções em Lingotamento Contínuo

Leandro Rodrigues Ramos[1], **Karin Satie Komati**[1],
Francisco de Assis Boldt[1], **Jefferson Oliveira Andrade**[1]

[1]Programa de Pós-graduação em Computação Aplicada (PPComp)
Campus Serra – Instituto Federal do Espírito Santo (IFES)

Resumo. *Obstruções das válvulas submersas no processo de lingotamento contínuo aumentam a frequência de interrupções na operação. Estas interrupções elevam o custo operacional, e podem provocar uma variedade de problemas de qualidade. A ausência de conjuntos de dados rotulados para as obstruções tem impedido a aplicação de métodos de aprendizado de máquina para predição desta anomalia no processo. Este trabalho buscou desenvolver técnicas semi-automáticas de rotulação de conjuntos de dados de referência. Como primeiro passo, aplicou-se uma técnica de clusterização sobre séries temporais fazendo uso do algoritmo DBSCAN. Os* clusters *gerados foram usados como sementes para um processo semi-supervisionado de propagação de rótulos. Este processo gerou uma base de dados que foi validada por especialistas e* 100% *dos dados rotulados como obstruções foram considerados corretamente rotulados.*

Abstract. *Clogging of submerged entry valves in the continuous casting process increase the frequency of interruptions in operation. These interruptions increase operating costs, and can cause a variety of quality problems. The absence of data sets labeled for clogging has prevented the application of machine learning methods for predicting this anomaly. This work sought to develop semiautomatic techniques for labeling reference data sets. As a first step, a clustering technique was applied over time series using the DBSCAN algorithm. The generated clusters were used as seeds for a semi-supervised label propagation process. This process generated a database that was validated by specialists and* 100% *of the data labeled as obstructions were considered correctly labeled.*

1. Introdução

Lingotamento contínuo é o processo pelo qual o metal fundido é solidificado em um produto semi-acabado, no caso deste trabalho em formato de placa [8]. O aço líquido é transferido do distribuidor para o molde por meio de um canal que é conhecido como *válvula submersa*. O aço é moldado e solidificado de maneira progressiva da superfície para o núcleo do veio (cada saída do lingotador). A obstrução de válvulas submersas se caracteriza como um dos problemas principais no lingotamento contínuo de aço.

As obstruções de válvulas submersas aumentam a frequência de interrupções do processo produtivo, para troca de válvulas, para troca de distribuidores, e até provocando uma parada completa da máquina. A injeção de gás argônio é uma técnica metalúrgica para formar uma cortina de gás que separa o fluxo do aço líquido da superfície refratária e que pode ser utilizada na prevenção e redução da obstrução. Porém todas estas ações

elevam o custo operacional, reduzem a produtividade da planta e podem provocar uma variedade de problemas de qualidade. O fluxo de aço líquido é controlado por um dispositivo de válvula gaveta, baseado no princípio de deslocamento paralelo de uma placa refratária, dotada de um orifício, entre duas outras, alinhando a abertura da placa móvel com os orifícios das placas finas. A obstrução muda os padrões de fluxo e as características dos jatos de aço que saem das válvulas, que podem interromper o fluxo no molde, levando a defeitos de superfície nos produtos de aço e até mesmo rompimentos (eventos conhecidos como *breakouts*). Os materiais que geram as obstruções também perturbam o fluxo, ficando presas no aço ou alterando a composição do mesmo, sendo que em ambos os casos originam defeitos [11].

A modelagem matemática deste problema vem sendo desenvolvida por diferentes abordagens. Yuan [15] usou equações hidrodinâmicas para a modelagem fenomenológica para o problema. Ometto [7] propôs um classificador baseado em árvores de decisão e *Gradient Boosting* para aproximar a relação não-linear entre a lista dos preditores e a variável alvo (obstrução), para tanto utilizou um conjunto de dados históricos (4 anos com aproximadamente 21.000 corridas de aço). Vannucci e Colla [13, 12] em seus trabalhos fazem uma combinação de técnicas clássicas envolvendo *perceptron* de múltiplas camadas (MLP do inglês *Multilayer Perceptron*), e árvores de decisão, dentre outras, objetivando detectar o problema. Para os modelos de classificação citados, a acurácia na detecção oscilou entre 74% e 80%. Variáveis estáticas do processo foram utilizadas como por exemplo, composição química do aço.

Uma corrida de lingotamento contínuo é uma sequência contínua de lingotamento, o aço presente no distribuidor origina diversas placas. A identificação do problema de obstrução de válvulas submersas a nível de corrida é importante para as equipes de operação, pois classifica se a corrida está propensa a ter obstrução ou não. Além disso, é relevante identificar em que trecho da placa lingotada ocorre o evento de obstrução, evitando assim a desclassificação de produtos de maneira inadequada. Buscando atender a ambos os requisitos, este trabalho apresenta técnicas para caracterização da obstrução através da análise de sinais e controles dinâmicos da linha, proveniente de sensores e indexados no tempo, ou seja, séries temporais multivariadas.

O artigo relata os resultados preliminares de um trabalho em evolução. Nesta etapa do trabalho busca-se desenvolver uma metodologia de rotulação de dados em séries temporais multivariadas. A rotulação é importante, pois não existe essa informação na base de dados da siderúrgica localizada na região Sudeste. Foram testados vários modelos estatísticos diferentes, visando identificar e agrupar anomalias nas séries temporais. Os grupos de dados "anômalos" foram validados manualmente por especialistas, e em seguida um processo de "transbordamento" de rótulos foi aplicado ao conjunto inicial de dados anômalos.

O estudo relatado neste artigo está claramente alinhado com o tema central do XL CSBC (Congresso da Sociedade Brasileira de Computação), "Artificialmente Humano ou Humanamente Artificial? Desafios para a Sociedade 5.0", visto que a técnica proposta se fundamenta em uma sinergia entre as capacidade humanas e as técnicas de aprendizado de máquina em um processo interativo que não seria efetivamente possível através do emprego de apenas um de qualquer dos dois elementos.

No que se refere à estrutura do artigo, na seção 2 são apresentados os principais conceitos e técnicas utilizados no trabalho. A metodologia empregada é descrita na seção 3, formando a base para a apresentação e análise dos resultados na seção 4. A seção 5 encerra o artigo com os comentários finais e conclusões.

2. Referencial Teórico

Nesta seção, descreve-se métodos de extração de características de forma sequencial em séries temporais. Destaca-se os métodos de aprendizado de máquina não supervisionados e técnicas semi-supervisionadas para propagação de rótulos (*label propagation*). O termo *semi-supervisionado* é utilizado neste trabalho para se referir a um processo interativo com etapas não supervisionadas, seguidas de etapas com supervisão manual.

2.1. Extração de Características em Séries Temporais

Uma série temporal é uma sequência de observações tomadas sequencialmente no tempo [1]. Uma série temporal univariada $X = [x_1, x_2, x_3, ..., x_T]$ representa uma sequência de medições da mesma variável (x) coletadas e indexadas ao longo do tempo. Séries temporais multivariadas são representadas por um conjunto $D = \{X_i\}_1^N$ de séries temporais univariadas (N=número de sinais), onde na melhor das hipóteses, possuem comprimento e taxas de amostragem iguais.

Para usar este conjunto como entrada para algoritmos de aprendizado de máquina supervisionados ou não supervisionados, cada série temporal X_i precisa ser mapeada em um espaço de características bem definidas de dimensionalidade M, por um vetor de características $\overrightarrow{x_i} = \{x_{i,1}, x_{i,2}..., x_{i,M}\}$. Em princípio, pode-se decidir mapear as séries temporais do conjunto D em uma matriz de N linhas e M colunas, escolhendo todos os M elementos de cada série temporal X_i como elementos do vetor $\overrightarrow{x_i}$. No entanto, outra abordagem de identificação de padrões é caracterizar as séries temporais em partes ou janelas, e para cada uma extrair propriedades de correlação, estacionariedade, entropia, dentre outros. Portanto, o vetor de características $\overrightarrow{x_i}$ pode ser construído aplicando métodos de extração de características $f_j : X_i \rightarrow x_{i,j}$ para as respectivas séries temporais X_i, resultando em um vetor de características $\overrightarrow{x_i} = (f_1(X_i), f_2(X_i), ..., f_M(X_i))$. Este vetor de características pode ser estendido com inclusão de atributos univariados $\{a_{i,1}, a_{i,2}, ...a_{i,U}\}_1^N$ [3].

2.2. Identificação de Anomalias por Método não Supervisionado

A clusterização é uma técnica de aprendizado não supervisionada que tem por objetivo identificar estruturas em um conjunto de dados não rotulados, organizando objetivamente os dados em grupos homogêneos, onde objetos de um grupo devem ser similares (ou relacionados) entre si, maximizando a dissimilaridade com objetos de outros grupos.

O uso desta técnica em séries temporais é uma atividade que vem sendo amplamente utilizada na comunidade de mineração de dados, no entanto, a maioria dos algoritmos executa a clusterização em toda a série temporal. Por outro lado, o agrupamento de subsequências em séries vem ganhando popularidade, sendo capaz de identificar *clusters* em subsequências de interesse em todo o fluxo de dados [9]. Define-se como uma subsequência de tamanho n em séries temporais $X = [x_1, x_2, x_3, ..., x_T]$ como $X_{i,n} = [x_i, x_{i+1}, ..., x_{i+n-1}]$, onde $1 \leq i \leq T - n + 1, n < T$.

A abordagem utilizada neste trabalho converte os dados brutos, presentes nas subsequências das séries, em vetores de características. Aplica-se então um algoritmo de clusterização DBSCAN [4], abreviação do termo (do inglês, *Density Based Spatial Clustering of Application with Noise*), sobre estes vetores no intuito de separar em *clusters* distintos as normas de operação e situações anômalas no processo. DBSCAN é um método de clusterização não paramétrico baseado em densidade, que é efetivo na identificação de *clusters* com formato arbitrário e de diferentes tamanhos. Sendo também capaz de identificar e separar os ruídos dos dados e detectar *clusters* e seus arranjos dentro do espaço de dados, sem qualquer informação preliminar sobre os grupos. A noção de *clusters* e o algoritmo DBSCAN se aplicam para qualquer espaço de características de alta dimensão. Os autores do método salientam ainda que a abordagem trabalha com qualquer função de distância, para este trabalho usamos duas métricas de distância: a distância Euclidiana e a distância de Mahalanobis. A distância de Mahalanobis leva em consideração o quanto um ponto está distante de sua distribuição (*clusters*), se mostrando efetiva na caracterização de *outliers (noise)*. É definida por:

$$D_M(p,q) = \sqrt{(p-q)^T \mathbf{C}^{-1}(p-q)} \tag{1}$$

onde $\mathbf{C}^{-1}$ é a inversa da matriz de covariância das variáveis independentes (q).

O DBSCAN é composto por dois parâmetros principais: $\varepsilon(eps)$ que representa a distância máxima entre dois pontos para que sejam considerados vizinhos, e *minPts* que representa o número mínimo de pontos que caracteriza uma região densa e consequentemente um *cluster*. Se for definido um valor baixo para *minPts*, aumenta-se a quantidade de *clusters* bem pequenos, no entanto, um valor muito grande pode impedir o algoritmo de criar *clusters*, terminando com uma base de dados composta apenas de anomalias.

Busca-se por regiões de alta densidade assinalando *clusters* às mesmas, ao passo que pontos em regiões menos densas não são sequer incluídos nos *clusters*, sendo rotulados como anomalias [2]. Entende-se como ponto, uma representação do espaço *n*-dimensional composto pelas características extraídas das subsequências nas séries temporais. Os *clusters* podem representar classes de operação normais (lingotamento normal, troca de panela, outros) e os *outliers* representar anomalias (lingotamento obstruído, trocas de válvula).

2.3. Método Semi-Supervisionado para Propagação de Rótulos

O conceito de propagação de rótulos (*label propagation*) foi introduzido por [16] como uma proposição para aprendizado de máquina semi-supervisionado. Dado um grafo ponderado finito $G = (V, E, W)$, formado por um conjunto de vértices V baseados em uma base de dados $X = \{x_i \mid i \in [1..n]\}$, um conjunto de arestas $E = (V \times V)$ e uma função de ponderação $w : E \to [0, 1]$. Se $w(i, j) > 0$, existe uma aresta entre x_i e x_j e $w(i, j)$ corresponde a uma medida de similaridade entre os mesmos [14]. Considerando ρ como uma métrica de distância definida no grafo, a matriz de similaridade w pode ser construída conforme a Equação (2), para alguma função h com decaimento exponencial no infinito, e.g., $h(x) = \exp(-x)$. Os pesos são controlados pelo parâmetro σ.

$$w(i,j) = h\left(\frac{\rho(x_i, x_j)^2}{\sigma}\right) \tag{2}$$

Uma matriz de transição probabilística para os rótulos pode ser definida através da normalização da matriz de similaridades conforme a Equação (3).

$$P(i,j) = \frac{w(i,j)}{\sum_{k \in V} w(i,k)} \tag{3}$$

Para este trabalho foi adotada uma abordagem de similaridade local [14] onde um grafo *KNN* correspondente é construído, onde somente as arestas entre os nós e seus vizinhos são ponderadas, gerando a matriz w' conforme a Equação (4). Com isto, gera-se a matriz *KNN* correspondente $\mathcal{P}$, conforme a Equação (5).

$$w'(i,j) = \begin{cases} w(i,j) & \text{se } x_j \in \textit{KNN}(x_i) \\ 0 & \text{caso contrario} \end{cases} \tag{4}$$

$$\mathcal{P}(i,j) = \frac{w'(i,j)}{\sum_{x_k \in \textit{KNN}(x_i)} w'(i,k)} \tag{5}$$

Algorithm 1 Label Propagation

Constrói a matriz de transição P_0 ▷ Conforme Eq. (3)
Inicializa os *labels*, $Y_0 \leftarrow [Y_0^l; -1]$ ▷ dados não rotulados $\leftarrow (-1)$
Calcula a matriz *KNN* $\mathcal{P}$ de P_0 ▷ Conforme Eq. (5)
Faça $t \leftarrow 0$
repeat
    $Y_{t+1} \leftarrow P_t \times Y_t$
    $Y_{t+1}^{(l)} \leftarrow Y_0^l$ ▷ *clamping*
    $P_{t+1} \leftarrow \mathcal{P}(P_t + \alpha Y_t Y_t^T)\mathcal{P}^T$
    Faça $t \leftarrow t + 1$
until Y_t convergir
Retorne Y_t

O algoritmo de propagação é executado para uma base de dados $X = \{X_l \cup X_u\}$ onde X_l representa os dados rotulados e X_u os dados não rotulados, $Y^{(l)}$ é a matriz resposta de rótulos. O algoritmo a cada iteração realiza um *clamping*, ou seja, reinicia os valores dos rótulos conhecidos. Um *fator de clamping* (α) pode ser utilizado para permitir flexibilização dos rótulos iniciais. Digamos que $\alpha = 0,2$, significa que serão retidos 80% da distribuição original dos rótulos. O Algoritmo 1, adaptado de [14], demonstra este procedimento.

3. Materiais e Métodos

Nesta seção será descrita a base de dados e detalhada a arquitetura geral do sistema, como as técnicas de clusterização e propagação de rótulos são usadas para a identificação de obstruções em lingotamento contínuo.

3.1. Base de Dados

A base de dados deste trabalho é proveniente de dados reais de uma empresa siderúrgica situada na região Sudeste. Os dados foram obtidos do processo siderúrgico de lingotamento contínuo e do refino do aço em convertedores a oxigênio. As variáveis independentes são representadas por séries temporais relevantes na caracterização do problema

e definidas pelos especialistas de processo das unidades técnicas de metalurgia, a Tabela 1 exemplifica algumas destas variáveis. Foram utilizados 2 meses de dados contínuos coletados de 10 em 10 segundos, o que corresponde a aproximadamente 500 corridas de lingotamento de aço. O intervalo da janela deslizante para extração de característica das séries temporais foi definido em 5 minutos. Considerando que velocidade nominal média de lingotamento é de 1m/min e que as placas produzidas possuem em média um comprimento de 11 metros, a janela de análise escolhida é satisfatória para detecção e classificação do problema e consequente julgamento das placas produzidas.

Tabela 1. Exemplos de variáveis dinâmicas do processo

Descrição	MIN	MAX	Unidade
Peso do carro distribuidor	0	70	ton
Velocidade do veio de lingotamento	0	2,5	m/min
Nível do molde	0	150	mm
Injeção de Argônio	0	60	NL/min
Abertura de válvula gaveta	0	100	%

O processo de extração de características e condicionamento de dados, aplicados aos 2 meses de dados do processo, originou uma base de dados de 27.251 amostras por 26 dimensões (características) considerando os 2 veios da máquina de lingotamento contínuo. As amostras possuem um identificador único no formato *{idVeio + nnnnnnn}* que representam o veio de lingotamento (3 ou 4) e o sequencial da janela deslizante respectivamente.

3.2. Modelagem

A Figura 1 apresenta um *pipeline* das técnicas abordadas neste trabalho com o intuito de se gerar rótulos válidos que caracterizem a obstrução e que servirão de suporte para a construção de futuros classificadores e modelos preditores para o problema.

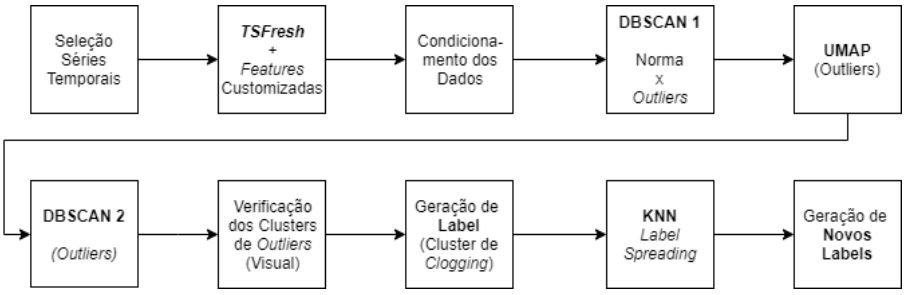

Figura 1. *Pipeline* para identificação de rótulos.

Para identificação do problema foi definido um conjunto de sinais do processo de lingotamento contínuo, representados por séries temporais multivariadas (bloco "Separação Séries Temporais"). Estas séries são divididas em subsequências de tempo das quais são extraídas um conjunto de características representativas dos sinais nos intervalos (bloco "***TSFresh*** + Features Customizadas"). Estes vetores de características passam por um processamento (bloco "Condicionamento dos Dados") e seguem em dois passos de clusterizações, combinados em diferentes espaços dimensionais fazendo uso do algoritmo DBSCAN (blocos "DBSCAN 1 Norma x Outliers" e "DBSCAN 2 (Outliers)"), obtendo-se os rótulos dos anomalias (bloco "Geração de label (Cluster de Clogging)").

Na sequencia, usa-se uma técnica de propagação de rótulos semi-supervisionada (bloco "KNN Label Spreading").

Para a modelagem as séries temporais foram divididas em janelas deslizantes (5 em 5 minutos), sendo gerados identificadores únicos para estes intervalos. Com base em análises exploratórias dos dados e entendimento da natureza física processo foram definidas características (exemplos na Tabela 2) que poderiam ser determinantes na separação realizada pelo método de clusterização, para diferenciar a normalidade do processo dos casos de pertubações que caracterizam a obstrução nas válvulas submersas. De forma complementar, adicionou-se novas características customizadas como a diferença entre mínimos e máximos nas janelas, e correlação entre a velocidade dos veios de lingotamento e abertura de válvula. Para a extração de características usa-se o pacote Python *TSFresh* (*Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests*) [3].

Tabela 2. Características selecionadas

Dicionário de Parâmetros	
Análise de abertura de válvula	mean_second_derivative_central
mean, median, minimum, maximum	variance_larger_than_standard_deviation
variance, standard_deviation	**Análise da válvula – 1ª derivada das séries**
absolute_sum_of_changes	count_above_mean, count_below_mean
linear_trend: [{'attr': 'slope'}]	variance_larger_than_standard_deviation
large_standard_deviation: [{'r': 0.5}]	number_crossing_m: [{'m': 0}]
longest_strike_above_mean	**Análise das variáveis de argônio**
longest_strike_below_mean	linear_trend: [{'attr': 'slope'}]
mean_change	

Atividades para condicionamento dos dados foram realizadas antes de se dar início no processo de geração dos *clusters*, tendo por objetivo eliminar ou minimizar a influência de situações como falta de dados e duplicidade de amostras [5]. Condições de parada de processo ($\sigma < 0,1$ em determinadas *features*) também foram filtradas. Características com variância zero foram eliminadas e mediu-se os coeficientes de correlação de Pearson entre as variáveis restantes. O coeficiente de correlação mensura o quão uma variável pode ser estimada ou explicada a partir de outra, assumindo valores na faixa entre -1 e +1. Os valores extremos da faixa indicam colinearidade perfeita, sendo -1 para correlação perfeita inversamente proporcional e +1 para correlação perfeita diretamente proporcional. Por fim, o coeficiente nulo indica independência estatística entre as variáveis. Variáveis com índice correlação superior a 0,98 foram eliminadas para redução de dimensionalidade do vetor de características das séries. Finalmente, visando estabelecer os mesmos graus de importância entre as variáveis independentes, os dados foram normalizados.

A primeira etapa de clusterização teve por objetivo separar a norma de operação dos *outliers*. Os hiper-parâmetros do DBSCAN foram ajustados para este propósito com valores de $\varepsilon = 1,35$ e $\varepsilon = 2$ com *minPts* correspondendo a 5% da amostra, fazendo uso das distâncias euclidiana e de Mahalanobis. Os parâmetros para os "*clusters* euclidianos" tiveram por objetivo caracterizar fortemente a norma de operação enquanto os parâmetros para Mahalanobis buscaram caracterizar os *outliers*. Estes *outliers* identificados foram projetados em um espaço dimensional reduzido fazendo uso do algoritmo UMAP (*Uniform Manifold Approximation and Projection*) [6]. UMAP é uma técnica de redução de dimensionalidade não linear (e não determinística) e preserva a natureza das relações entre os pontos após a projeção no espaço dimensional reduzido.

Uma segunda etapa de clusterização é realizada sobre o espaço projetado pelo UMAP, considerando apenas o universo de pontos classificados como *outliers* na primeira etapa. Para esta fase foi realizado uma calibração dos hiper-parâmetros do DBSCAN considerando o *sillhouete score* [10] como métrica de avaliação da qualidade dos *clusters*. Uma análise visual de amostras dos elementos destes *clusters* proporciona uma rápida identificação dos principais casos de interesse, gerando um conjunto inicial de rótulos que serve de base para o passo final deste trabalho que envolve a propagação de rótulos. O fator de *clamping* escolhido é $\alpha = 20\%$.

4. Resultados

Todas as amostras da base de dados foram submetidas ao algoritmo de clusterização (DBSCAN) com dois objetivos distintos:

1. Caracterização da norma: Uso de distância euclidiana e hiper-parâmetros (ε, *minPts*) ajustados para caracterizar a norma de operação e aumentar a zona de fronteira com os *outliers*.
2. Seleção de *outliers*: Uso da distância de Mahalanobis e hiper-parâmetros ajustados para segregar de forma mais efetiva os *outliers*.

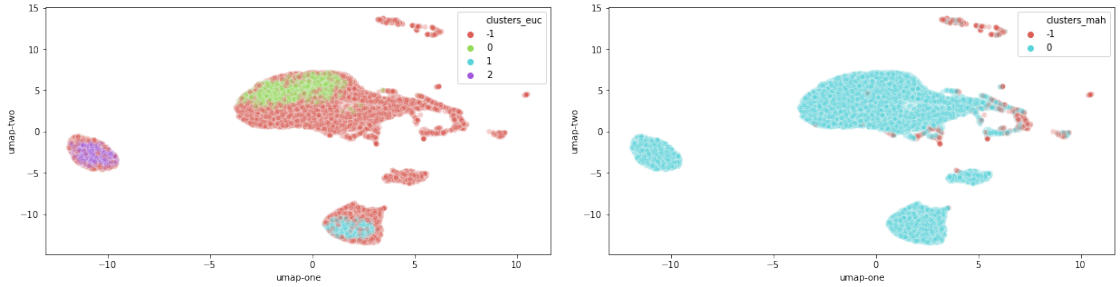

Figura 2. Clusterização inicial (*Outliers* x Norma).

Como resultado foram geradas 9.636 amostras de norma e 1.258 anomalias. O gráfico à esquerda na Figura 2 apresenta uma visualização UMAP dos resultados contendo os *clusters* euclidianos da norma $(0, 1, 2)$ e os *outliers* (-1) (em vermelho) e o gráfico à direita na Figura 2 gerados com a distância de Mahalanobis. Com as anomalias caracterizadas, evolui-se no *pipeline* para uma segunda etapa de clusterização. As anomalias são projetadas em espaço dimensional reduzido (UMAP) e novamente executa-se o DBSCAN, tendo os seus hiper-parâmetros calibrados e avaliados por *sillhouete score*. Nesta etapa busca-se uma caracterização ainda maior de cenários anômalos distintos no processo (obstrução, troca de válvula, saída e retorno de processo, outros). Os *clusters* gerados (Figura 3) serviram de base para uma análise visual e geração das "sementes" que foram submetidas ao processo subsequente de propagação de rótulos.

A etapa final consiste na aplicação de uma técnica semi-supervisionada de propagação de rótulos. Neste contexto foram definidas 3 classes distintas a serem utilizadas pelo algoritmo. A classe de operação normal (rótulo 0) foi populada com as 9.636 amostras dos *clusters* de norma geradas na 1ª etapa de clusterização. Uma segunda classe, denominada "anomalias conhecidas" (rótulo 3) foi populada, contendo os *clusters* 4 e 7 (em sua totalidade) provenientes da 2ª etapa de clusterização. Esta classe foi bem caracterizada na etapa anterior e representa cenários anômalos (por exemplo, a troca de

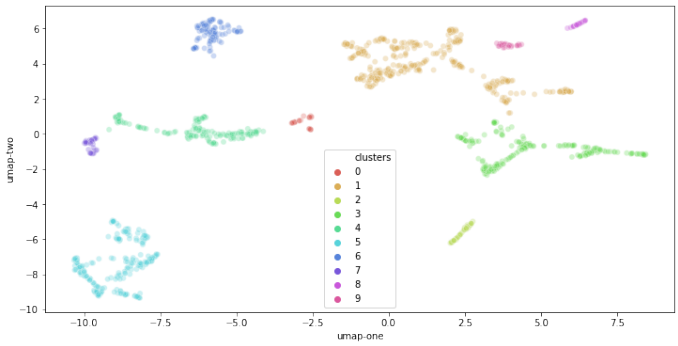

Figura 3. Clusterização dos *outliers*.

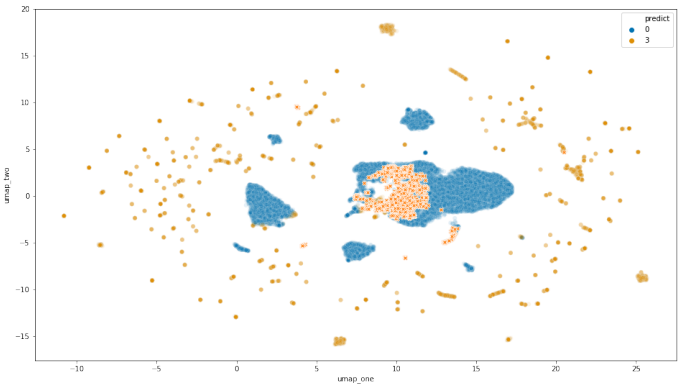

Figura 4. Propagação de rótulos: Norma (0), Anomalias Conhecidas (3), Obstruções (x).

válvulas) diferentes do problema alvo de obstrução. Adicionalmente foram acrescentadas a esta classe todos os intervalos de parada de processo que foram filtrados durante a fase de extração de características, totalizando 6.984 amostras. Por fim, para popular a classe de obstrução (rótulo 7), foram cruzados as amostras de *outliers* identificadas pelo método de clusterização com macro-intervalos de prováveis obstruções sugeridos pela metalurgia e apontados (de uma forma indireta) pelos sistemas de qualidade da empresa. Deste cruzamento, pode-se observar 152 amostras consistentes de obstrução que serviram para semear o processo de propagação de rótulos. As demais amostras (17.182) foram consideradas não rotuladas (rótulo -1) e consistem o espaço de propagação do algoritmo. Inicia-se então o processo semi-supervisionado com as seguintes etapas:

1. Remarcação dos rótulos da norma.
2. Execução do algoritmo de propagação de label: *Kernel* kNN, $n\_neighbors = 7$, $\alpha = 20\%$.
3. Análise visual dos resultados.
4. População das listas de exclusões (falso-positivos).
5. Reinicia passo 1 até convergir.

A Figura 4 apresenta o cenário final de convergência em espaço UMAP. Totalizou 24.454 amostras de norma, 7.566 amostras de anomalias conhecidas e 1.934 de obstruções (5,7%). Deste total deve-se desconsiderar 202 casos de obstruções e 138 casos de anomalias conhecidas, presentes nas listas de exclusões e que foram populadas por observação durante o processo de propagação. Vale ressaltar que para um modelo de classificação

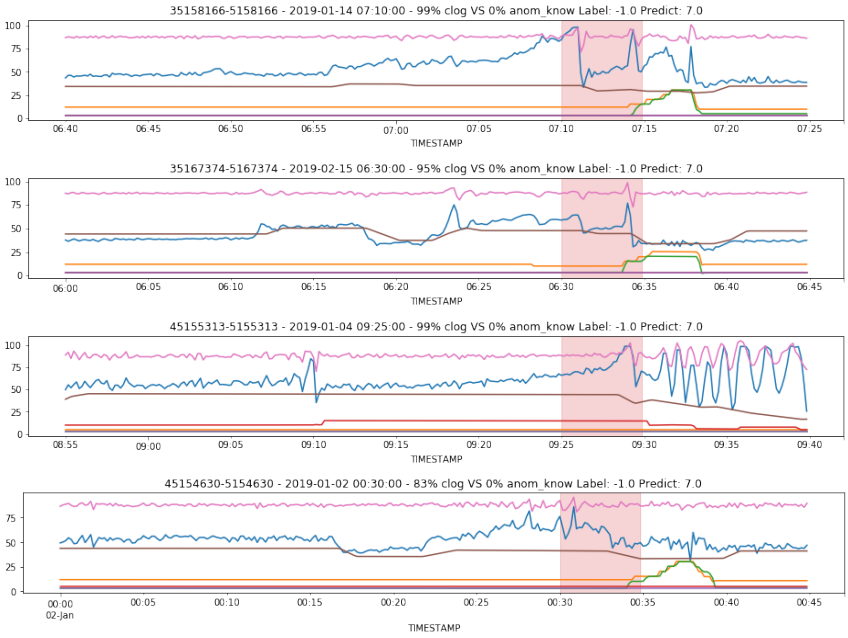

Figura 5. Identificação de Obstruções.

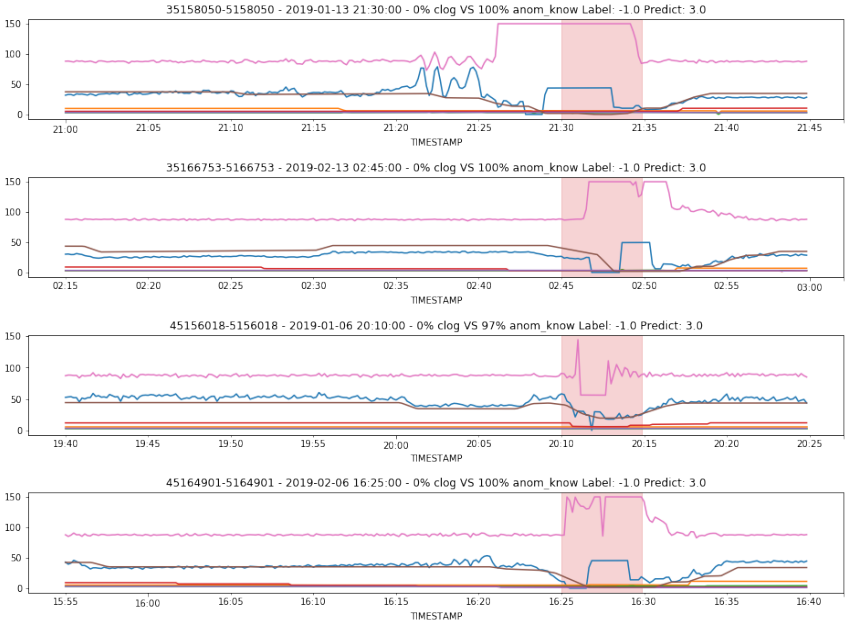

Figura 6. Identificação de Anomalias Conhecidas.

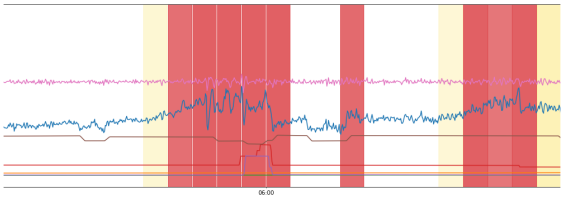

Figura 7. Probabilidades de obstrução sequenciada no tempo.

futuro, estas exclusões precisarão ser devidamente rotuladas.

Visando acelerar a análise exploratória e consequente detecção das anomalias, foi construída uma ferramenta em *Python* para visualização dos resultados. Gera-se visões resultantes de cruzamento dos *ids* das amostras rotuladas, matriz probabilística do kNN e os dados brutos das séries temporais originais. As Figuras 5 e 6 apresentam exemplos de casos identificados (e confirmados) através do uso deste ferramental disponibilizado. De forma complementar, "mapas de calor" associados às probabilidades de anomalia foram gerados sobre as séries temporais (Figura 7).

Deve-se ressaltar que estes resultados foram analisados por especialistas do domínio de interesse que confirmaram os resultados obtidos como correspondendo a eventos de obstrução reais. Ou seja, ao final do processo semi-supervisionado de propagação de rótulos, 100% dos eventos rotulados como obstruções foram validados.

5. Conclusões

A obstrução em máquinas de lingotamento contínuo é um fenômeno de difícil identificação. O contexto industrial avaliado não possuía o evento caracterizado de forma direta, seja nos sistemas de produção ou nos historiadores do processo. Este trabalho trouxe uma técnica capaz de indicar regiões de alta probabilidade para o problema de obstrução bem como regiões contendo outras anomalias conhecidas, o que também se mostrou útil para as equipes de metalurgia. O método proposto acelera significativamente a geração de rótulos visando popular uma base de treinamento para modelos preditores, e se mostrou robusto na detecção de anomalias em veios de lingotamento com comportamentos operacionais diferentes, indicativo de que o mesmo pode ser generalizado para diferentes máquinas.

De posse da base rotulada, o próximo passo envolve a construção de futuros modelos classificadores visando a predição destas anomalias em tempo de processo, como suporte à decisão para julgamento da qualidade dos produtos produzidos.

Referências

[1] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung. *Time series analysis: forecasting and control*. Wiley, 2016.

[2] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Computing Surveys*, 41(3):1–58, jul 2009.

[3] M. Christ, N. Braun, J. Neuffer, and A. W. Kempa-Liehr. Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (TSFresh – A Python package). *Neurocomputing*, 307:72–77, sep 2018.

[4] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *2nd International Conference on Knowledge Discovery and Data Mining*, pages 226–231, 1996.

[5] S. García, J. Luengo, and F. Herrera. Tutorial on practical tips of the most influential data preprocessing algorithms in data mining. *Knowledge-Based Systems*, 98:1–29, apr 2016.

[6] L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[7] L. Ometto, S. Challapalli, M. Polo, G. Cestari, A. Villagrossi, M. Sandri, and E. Pellegrini. Successful Use Case Applications of Artificial Intelligence in the Steel Industry. In *AISTech2019 Proceedings of the Iron and Steel Technology Conference*, pages 2573–2584. AIST, 2019.

[8] J. J. M. Peixoto. Modelamento físico e matemático do fluxo no interior de um molde de lingotamento contínuo de beam blank alimentado com duas válvulas submersas tubulares. Master's thesis, Programa de Pós-Graduação em Engenharia de Materiais. Escola de Minas, Universidade Federal de Ouro Preto., 2016.

[9] S. Rodpongpun, V. Niennattrakul, and A. Ratanamahatana. Selective Subsequence Time Series clustering. *Knowledge-Based Systems*, 35:361–368, 2012.

[10] P. J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53 – 65, 1987.

[11] B. G. Thomas and H. Bai. Tundish Nozzle Clogging-Application Of Computational Models. In *18rd Process Technology Division Conference Proceedings*, volume 18. Iron and Steel Society, 2001.

[12] M. Vannucci and V. Colla. Novel classification method for sensitive problems and uneven datasets based on neural networks and fuzzy logic. *Applied Soft Computing Journal*, 11(2):2383–2390, 2011.

[13] M. Vannucci, V. Colla, G. Nastasi, and N. Matarese. Detection of rare events within industrial datasets by means of data resampling and specific algorithms. *International Journal of Simulation: Systems, Science and Technology*, 11(3):1–11, 2010.

[14] B. Wang, Z. Tu, and J. Tsotsos. Dynamic label propagation for semi-supervised multi-class multi-label classification. *Proceedings of the IEEE International Conference on Computer Vision*, 52:425–432, 12 2013.

[15] F. Yuan, X. Wang, J. Zhang, and L. Zhang. Online forecasting model of tundish nozzle clogging. *Journal of University of Science and Technology Beijing: Mineral Metallurgy Materials (Eng Ed)*, 13(1):21–24, feb 2006.

[16] X. Zhu and Z. Ghahramani. Learning from labeled and unlabeled data with label propagation. Technical Report CMU-CALD-02-107, Carnegie Mellon University, Pittsburgh, PA, 2002.