ConBRepro
X CONGRESSO BRASILEIRO DE ENGENHARIA DE PRODUÇÃO
EVENTO ON-LINE
02 a 04 de dezembro 2020

Modelagem de um Processo Produtivo utilizando Técnicas de Aprendizado de Máquina

Simone Massulini Acosta
Departamento de Eletrônica – UTFPR Câmpus Curitiba
Anderson Levati Amoroso
Departamento de Eletrônica – UTFPR Câmpus Curitiba

Resumo: Em um processo industrial pode ser definido um conjunto de causas que produzem determinado efeito sobre uma ou mais características da qualidade de um produto. Como resultado pode-se ter a produção de produtos não conformes às especificações, que podem ser mensurados através da fração de produtos não conformes. A modelagem da fração de produtos não conformes pode ser realizada utilizando-se diferentes técnicas e modelos de regressão, sendo as técnicas de aprendizado de máquina muito utilizadas para a modelagem de dados de processos. O objetivo deste artigo foi modelar a fração de produtos não conformes às especificações de uma indústria utilizando as técnicas de aprendizado de máquina: Redes Neurais Artificiais, Regressão por Vetores de Suporte, Regressão por Vetores de Relevância, Regressão por Processo Gaussiano, Árvores de Regressão, Árvores de Modelos, Floresta Aleatória, K-NN e *Random Vector Functional Link*. Através dos resultados verifica-se que o modelo RVR apresenta melhor desempenho no ajuste do modelo aos dados do processo do que os demais modelos analisados. Através dos resultados pode-se considerar que todos os modelos baseados nas técnicas de aprendizado de máquina representam adequadamente a fração de produtos não conformes às especificações do processo produtivo.

Palavras-chave: Modelagem de dados, aprendizado de máquina, fração de produtos não conformes.

Modeling a Productive Process based on Machine Learning Techniques

Abstract: In an industrial process, a set of causes can be defined that produce a specific effect on one or more characteristics of the quality of a product. As a result, it is possible to have products that do not conform to specifications, measured by the nonconforming fraction. The modeling of the nonconforming fraction can be performed using different techniques and regression models. Machine learning techniques are widely used for modeling process data. The purpose of this article was to model the nonconforming fraction of an industry using machine learning techniques: Artificial Neural Networks, Support Vector Regression, Relevance Vector Regression, Gaussian Process Regression, Regression Trees, Model Trees, Random Forest, K-NN and Random Vector Functional Link. The results show that the RVR model performs better in adjusting the model to the process. The results of this study indicate that all models based on machine learning techniques are adequate tools for prediction the nonconforming fraction in the production process.

Keywords: Data modeling, machine learning, nonconforming fraction

1. Introdução

Em um processo industrial, muitas vezes não se consegue controlar todas as causas de variação que produzem determinado efeito sobre as características da qualidade dos produtos, pois certas causas são inerentes ao processo. As causas de variação que podem ser controladas podem interferir em um processo gerando produtos com características da qualidade não conformes às especificações preestabelecidas, que podem ser mensuradas através da fração de produtos não conformes (MONTGOMERY, 2004).

A fração não conforme é definida como a razão entre o número de unidades não conformes da amostra e o tamanho da amostra (MONTGOMERY, 2004). A fração não conforme pode compreender a razão entre dois números discretos, denominada de percentual e comumente modelada como distribuição Binomial, ou a razão entre dois números contínuos, denominada de proporção e que pode ser modelada como distribuição Beta. As características da qualidade do tipo fração possuem as observações expressas no intervalo [0,1].

A análise de regressão consiste na modelagem da relação entre as características da qualidade e as variáveis de controle do processo. Conforme Montgomery e Runger (2003), um modelo de regressão que apresenta um bom ajuste usualmente permite gerar boas estimativas dos efeitos dos fatores, pois é possível prever a fração de produtos não conformes em função do ajuste das variáveis do processo.

Modelos de regressão baseados em técnicas de aprendizado de máquina têm sido propostos na literatura. O aprendizado de máquina trata do desenvolvimento de técnicas computacionais sobre o aprendizado e a construção de sistemas capazes de aprender e melhorar seu desempenho baseado em experiências acumuladas através da solução de problemas anteriores (MITCHELL, 1997).

O objetivo deste artigo foi modelar a fração de produtos não conformes às especificações de uma indústria curtidora de couro utilizando as técnicas de aprendizado de máquina: Redes Neurais Artificiais, Regressão por Vetores de Suporte, Regressão por Vetores de Relevância, Regressão por Processo Gaussiano, Árvores de Regressão, Árvores de Modelos, Floresta Aleatória, K-NN e *Random Vector Functional Link*.

2. Referencial teórico

Nesta sessão apresenta-se uma breve descrição dos referenciais teóricos das técnicas utilizadas para o desenvolvimento dos modelos de regressão deste artigo.

2.1. Redes neurais artificiais

Segundo Haykin (2009), a principal propriedade de uma Rede Neural Artificial (RNA) é a sua habilidade de aprender a partir de seu ambiente e de melhorar o seu desempenho através do aprendizado. Entre os principais tipos de RNAs com a arquitetura *feedforward* de camadas múltiplas está o Perceptron Multicamadas (*Multilayer Perceptron*, MLP). Conforme Braga *et al*. (2012) para as redes MLP pode ser utilizado o algoritmo *error backpropagation*. Existem várias modificações do algoritmo *backpropagation* que visam melhorar seu desempenho, sendo uma destas modificações a Levenberg-Marquardt. Para mais informações sobre RNAs sugere-se consultar Vapnik (1998), Haykin (2009) e Faceli et al. (2011).

2.2. Regressão por vetores de suporte

As Máquinas de Vetores de Suporte (*Support Vector Macchine*, SVM) foram originalmente desenvolvida para classificação, sendo estendidas para problemas de regressão e

denominadas Regressão por Vetores de Suporte (*Support Vector Regression*, SVR) (VAPNIK, 1998).

Na SVR tem-se dois conceitos importantes, que são o ε-tubo e a função de perda ε-insensível, que ignora erros que estão além de uma certa distância dos valores considerados válidos. As amostras fora do ε-tubo correspondem aos vetores de suporte (KECMAN, 2001).

Quando um modelo não linear é necessário utiliza-se uma função kernel, que possibilita que os dados de entrada originais sejam mapeados no espaço de características de elevada dimensão. Utilizando o método do multiplicador de Lagrange, equações (1) e (2), pode-se obter a função objetivo, equação (3) (SCHÖLKOPF; SMOLA, 2004).

$$\underset{\alpha,\alpha}{\text{Maximizar}} \; -\frac{1}{2}\sum_{i,j=1}^{n}\left(\alpha_i-\alpha_i^*\right)\left(\alpha_j-\alpha_i^*\right)k\left(x_i,x_j\right)-\varepsilon\sum_{i=1}^{n}\left(\alpha_i+\alpha_i^*\right)+\sum_{i=1}^{n}y_i\left(\alpha_i-\alpha_i^*\right) \tag{1}$$

$$\text{com as restrições}\begin{cases}\sum_{i=1}^{n}\left(\alpha_i-\alpha_i^*\right)=0\\ \alpha_i,\alpha_i^*\in[0,C]\end{cases} \tag{2}$$

$$f\left(x\right)=\sum_{i=1}^{n}\left(\alpha_i-\alpha_i^*\right)k\left(x_i,x_j\right)+b \tag{3}$$

onde α_i e α_i^* representam as variáveis de Lagrange, $k\left(x_i,x_j\right)$ é a função kernel, *b* o termo de limiar e *C* é a constante de regularização.

Neste artigo foi utilizado o kernel Função de Base Radial (*Radial Basis Function*, RBF), equação (4), onde $\gamma=1/2\nu^2$ e $\nu>0$ é o parâmetro que define a largura do kernel.

$$k\left(x_i,x_j\right)=exp\left(-\frac{1}{2\nu^2}\left\|x_i-x_j\right\|^2\right)=exp\left(-\gamma\left\|x_i-x_j\right\|^2\right) \tag{4}$$

A capacidade de generalização do modelo SVR depende da seleção dos parâmetros *C* e ε e do parâmetro do kernel RBF durante a etapa de treinamento. Das diversas propostas de seleção de parâmetros, a busca em grade e a validação cruzada são as mais utilizadas (FACELI *et al*., 2011). Para detalhes sobre SVMs recomenda-se consultar Vapnik (1998), Kecman (2001) e Schölkopf e Smola (2004).

2.3. Regressão por vetores de relevância

As Máquinas de Vetores de Relevância (*Relevance Vector Machine*, RVM) são modelos probabilísticos baseados em métodos kernel com forma funcional similar as SVMs (TIPPING, 2000). A RVM utilizada para problemas de regressão é denominada de Regressão por Vetores de Relevância (*Relevance Vector Regression*, RVR).

A RVR adota uma estrutura totalmente probabilística e utiliza a probabilidade a priori sobre os pesos (ω_r) do modelo regidos por um conjunto de hiperparâmetros (α_r), um associado a cada peso, cujos valores mais prováveis são estimados de forma iterativa a partir dos dados de treinamento (TIPPING, 2001). Após o processo de otimização, as entradas correspondentes aos pesos remanescentes não nulos são denominados vetores de relevância (BISHOP, 2006).

Na RVR é necessário selecionar somente os parâmetros da função kernel durante a etapa de treinamento, sendo os demais parâmetros estimados automaticamente pelo procedimento de aprendizado. Neste artigo foi utilizado o kernel RBF, apresentado na

equação (4). Para mais informações sobre RVMs sugere-se consultar Tipping (2000), Tipping (2001) e Bishop (2006).

2.4. Regressão por processo gaussiano

A Regressão por Processo Gaussiano (*Gauss Process Regression*, GPR) consiste em estimar uma função para entradas arbitrárias, dado um conjunto a priori de treinamento (MACKAY, 1998). Um processo Gaussiano (*Gauss Process*, GP) é um método Bayesiano de regressão que consiste em uma extensão da distribuição Gaussiana multivariada (RASMUSSEN; WILLIAMS, 2006).

O GP é especificado por sua média a priori e sua função de covariância, também denominada kernel, que determina a estrutura das funções no espaço de funções no processo Gaussiano. Schulz *et al.* (2018) sugerem utilizar como função de covariância o kernel RBF, apresentado na equação (4). Para mais informações sobre GPR, consultar MacKay (1998) e Rasmussen e Williams (2006).

2.5. Árvore de regressão

Uma Árvore de Regressão (*Regression Tree*, RT) é uma técnica em que os dados do grupo de treinamento são particionados em subespaços menores que destaquem características que melhor funcionem para prever a variável objetivo. Para prever o valor da resposta de uma observação verifica-se a região a qual a resposta pertence e calcula-se a média dos valores da variável resposta das amostras do conjunto de treinamento pertencentes à região analisada (SAFAVIAN; LANDGREBE, 1991).

O processo de criação de uma RT é dividido em duas etapas: criação de uma árvore completa, e a poda da árvore criada, com o objetivo de evitar o *overfitting* (sobreajuste). Uma árvore é construída por particionamento recursivos no espaço das covariáveis sendo que cada particionamento recebe o nome de nó e cada resultado final recebe o nome de folha (IZBICKI; SANTOS, 2019). Para mais detalhes sobre RT recomenda-se consultar Safavian e Landgrebe (1991) e Elith *et al*. (2008).

2.6. Árvore de modelos

Uma Árvore de Modelos (*Model Tree*, MT) utiliza particionamento recursivo para construir um modelo linear por partes na forma de uma árvore de modelo (Quinlan, 1992). A ideia é dividir os casos de treinamento da mesma maneira que ocorre nas árvores de decisão, usando um critério de minimização da variação dos valores nos subconjuntos em vez de maximizar o ganho de informação. M5 (Quinlan, 1992) constrói modelos baseados em árvore porém, enquanto as árvores de regressão (Breiman *et al*., 1984) possuem valores em suas folhas, as árvores construídas por M5 podem ter modelos lineares multivariados. Neste trabalho, foi utilizado o modelo baseado em regras M5 com *boosting* e correções baseadas nos vizinhos mais próximos ao conjunto de dados de treinamento (Quinlan, 1993; Fernández-Delgado *et al*., 2019). Para mais detalhes sobre MT recomenda-se a leitura de Quinlan (1992), Quinlan (1993) e Fernández-Delgado *et al*. (2019).

2.7. Floresta aleatória

Uma Floresta Aleatória (*Random Forest*, RF) é uma combinação de árvores preditoras em que cada árvore depende dos valores de vetores aleatórios amostrados de forma independente e distribuídos igualmente para todas as árvores na floresta (BREIMAN, 2001). Na RF, cada nó é dividido usando o melhor entre um subconjunto de preditores escolhidos aleatoriamente naquele nó (LIAW; WIENER, 2002).

O algoritmo de RF possui dois parâmetros, que são o número de árvores na floresta e o número de atributos selecionados para determinar a divisão em cada nó das árvores. De acordo com Probst e Boulesteix (2018), o número de árvores na floresta deve ser grande de modo que cada característica candidata tenha oportunidades suficientes para ser

selecionada. Na prática, o desempenho atinge um patamar com algumas centenas de árvores para a maioria dos conjuntos de dados. Para mais detalhes sobre RF recomenda-se a leitura de Breiman (2001) e Liaw e Wiener (2002).

2.8. K-vizinhos mais próximos

No método conhecido como K-vizinhos mais próximos (*K-nearest neighbours*, K-NN) não existe a fase de aprendizagem com a construção de modelo e posterior generalização para todos os dados, mas sim a memorização dos dados de treinamento (FACELI et al., 2011).

O método K-NN tem como base estimar a função de regressão para uma dada configuração das covariáveis X com base nas variáveis respostas Y dos k-vizinhos mais próximos a X (BENEDETTI, 1977).

A função de regressão avaliada em X é estimada utilizando-se uma média local das respostas dos k vizinhos mais próximos a X no espaço das covariáveis. O parâmetro k pode ser selecionado através da validação cruzada (IZBICKI; SANTOS, 2019). Para mais detalhes sobre K-NN recomenda-se consultar Benedetti (1977) e Haykin (2009).

2.9. *Random Vector Functional Link*

Pao *et al*. (1994) propuseram a rede neural *Random Vector Functional Link* (RVFL). A RVFL é uma extensão da rede neural *feedforward* de camada simples (*Single Layer Feedforward Neural*, SLFN) com conexões diretas adicionais da camada de entrada para a camada de saída (Qiu et al., 2018). A rede RVFL possui um conjunto de nós chamados nós de realce, que são equivalentes aos neurônios da camada oculta da SLFN. Os valores dos pesos da camada de entrada para a camada oculta na RVFL são gerados aleatoriamente em um domínio adequado e mantidos fixos na etapa de aprendizagem (Zhang e Suganthan, 2015). Para mais detalhes sobre RVFL recomenda-se a leitura de Qiu *et al*. (2018) e Zhang e Suganthan (2015).

3. Procedimentos para o desenvolvimento dos modelos e análise de desempenho

Normalmente, o procedimento para o desenvolvimento dos modelos de regressão segue os seguintes passos: obtenção dos dados, análise e preparação dos dados, seleção das variáveis, normalização e divisão dos dados, escolha e seleção dos melhores parâmetros da técnica utilizada, simulações de treinamento e de teste e análise do desempenho.

Para análise do desempenho podem ser utilizadas estratégias de minimização do erro residual. O erro residual (e_t) é a diferença entre o valor real (y_i) e o valor estimado pelo modelo (y_t), equação (5). A equação (6) apresenta o erro médio absoluto (*Mean Absolute Error*, MAE), a equação (7) apresenta o erro médio quadrático (*Mean Squared Error*, MSE) e na equação (8) é apresentada a raiz do erro médio quadrático (*Root Mean Squared Error*, RMSE), sendo *n* o número de amostras.

$$e_t = y_i - y_t \tag{5}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - y_t| = \frac{1}{n}\sum_{i=1}^{n}|e_t| \tag{6}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_t)^2 = \frac{1}{n}\sum_{i=1}^{n}(e_t)^2 \tag{7}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - y_t)^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(e_t)^2} \tag{8}$$

Os resíduos dos modelos de regressão devem ser analisados com relação a normalidade, independência e variância. A verificação da normalidade dos resíduos pode ser verificada

por meio do teste de normalidade de Shapiro-Wilk, Anderson-Darling, dentre outros. A estatística de Durbin-Watson pode ser utilizada para verificação da hipótese de que os resíduos são independentes (MONTGOMERY; RUNGER, 2003).

A verificação da homoscedasticidade, ou seja, se os resíduos possuem variância aproximadamente constante, pode ser analisada pelo teste de Levene e pelo gráfico dos resíduos versus os valores estimados. Os resíduos devem ser aleatórios e não devem seguir padrões específicos, conforme os apresentados em Montgomery e Runger (2003). Conforme Werkema e Aguiar (1996), pode-se utilizar o gráfico dos resíduos padronizados versus os valores estimados e, se os resíduos padronizados são i.i.d., em torno de 95% dos resíduos devem estar no intervalo [-2,2] e 99% no intervalo [3;+3].

O resíduo padronizado (e_{pt}) utiliza a variância estimada do resíduo ($\hat{\sigma}^2$), conforme apresentado na equação (9), onde n representa o número de observações e h o número de variáveis de controle

$$e_{pt} = \frac{(y_i - y_t)}{\sqrt{\hat{\sigma}^2}} = \frac{(y_i - y_t)}{\sqrt{\dfrac{\sum_{t=1}^{n}(y_i - y_t)^2}{n-h-1}}} = \frac{e_i}{\sqrt{\dfrac{\sum_{t=1}^{n} e_i^{\,2}}{n-h-1}}} \tag{9}$$

O modelo de regressão validado pode ser utilizado para o monitoramento do processo através da carta de controle dos resíduos padronizados, com a definição dos limites de controle e a verificação se o processo está sob controle. As estimativas de média ($\bar{e}_{pt}$) e do desvio padrão ($\sigma_{e_{pt}}$) dos resíduos padronizados dos dados do processo sob controle são utilizadas para calcular o limite superior de controle (LSC), o limite inferior de controle (LIC) e a linha central (LM) da carta de controle univariada de Shewhart, conforme equações (10), (11) e (12)

$$LSC = \bar{e}_{pt} + \lambda\sigma_{e_{pt}} \tag{10}$$

$$LM = \bar{e}_{pt} \tag{11}$$

$$LIC = \bar{e}_{pt} - \lambda\sigma_{e_{pt}} \tag{12}$$

onde λ é a constante que define a largura dos limites de controle correspondente a uma região de controle (1-α) e um número médio desejado de amostras até um alarme falso (NMA_0). Habitualmente, utiliza-se o valor de λ igual a 3 (três) devido à aproximação pela distribuição Normal, correspondendo a NMA_0 = 370,4.

4. Estudo aplicado

Neste artigo foi desenvolvida a modelagem da fração de produtos não conformes às especificações de uma empresa curtidora de couro, produtora de couro acabado e fornecedora para as indústrias de calçados e artefatos em couro.

A etapa *wet blue* do processo produtivo avaliado consiste em: o classificador recebe um lote de diferentes tamanhos contendo as matérias-primas e verifica se as características de qualidade satisfazem às especificações, por métodos cognitivos. As matérias-primas que não satisfazem às especificações são classificadas como produtos não conformes e a fração de produtos não conformes às especificações, por lote, é considerada a variável dependente (característica da qualidade). Os dados coletados contemplaram uma amostra de 713 lotes, sendo que a fração de produtos não conformes do processo segue a distribuição Beta.

Os fatores controláveis definidos como variáveis independentes para a modelagem da fração de produtos não conformes foram: a seleção da matéria prima conforme qualidade e preço (com cinco níveis diferentes), a procedência da matéria-prima adquirida pela empresa (com cinco níveis), o classificador que inspeciona as matérias-primas (com três níveis) e o estado de rebaixamento da matéria-prima (com dois níveis)

Para a construção dos modelos de regressão as variáveis independentes qualitativas seleção, procedência, classificador e rebaixamento foram substituídas pelas variáveis *dummy*. As novas variáveis independentes foram definidas como: seleção tipo 2 (x_1), seleção tipo 3 (x_2), seleção tipo 4 (x_3), seleção tipo 5 (x_4), procedência 2 (x_5), procedência 3 (x_6), procedência 4 (x_7), procedência 5 (x_8), classificador 2 (x_9), classificador 3 (x_{10}) e rebaixamento (x_{11}).

Para a definição das variáveis independentes relevantes para a construção dos modelos de regressão foi realizado o teste de correlação de Pearson, sendo selecionadas as variáveis seleção (tipo 2, tipo 3, tipo 4 e tipo 5), classificador (2 e 3) e rebaixamento como significativas estatisticamente para explicar a variável dependente fração de produtos não conformes.

Após a análise e preparação dos dados o conjunto total de dados foi dividido aleatoriamente em dois subconjuntos: 70% (499 observações) para o grupo de treinamento e 30% (214 observações) para o grupo de teste. Os cálculos e as simulações necessárias foram desenvolvidas com o programa R®.

A arquitetura selecionada para a RNA foi a MLP com uma camada oculta. Os parâmetros selecionados para a RNA foram: algoritmo de treinamento Levenberg-Marquardt, função de ativação logística para a camada oculta e linear para a camada de saída, taxa de aprendizado de 0,01 e 19 neurônios na camada oculta. O método de seleção dos parâmetros do modelo SVR foi a busca em grade em conjunto com a validação cruzada (10-*fold*) nos dados de treinamento. O espaço de busca para os parâmetros foi: $C \in [1;50]$, $\varepsilon \in [0,001;1]$ e $\gamma \in [0,001;1]$. Os melhores valores obtidos para os parâmetros foram: C = 10, ε = 0,03 e γ = 0,2.

Para a seleção do parâmetro do kernel RBF do modelo RVR o espaço de busca foi $\gamma \in [0,001;1]$. O método de seleção utilizado foi a validação cruzada (10-*fold*) e o melhor valor obtido para o parâmetro foi γ = 0,2127. Para o modelo GPR, para a seleção do parâmetro do kernel RBF foi utilizado o espaço de busca $\gamma \in [0,001;1]$. O método de seleção utilizado foi a validação cruzada (10-*fold*), sendo que o melhor valor obtido para o parâmetro foi γ = 0,29306.

Para a Árvore de Regressão (RT), para a obtenção da melhor partição em cada etapa do processo de criação foi utilizado o método da validação cruzada. Os parâmetros ajustáveis da Árvore de Modelos (MT) foram selecionados pelo método da validação cruzada. O número de comitês de treinamento selecionado foi 2 e o número de vizinhos mais próximos foi 9. Para a RF os parâmetros selecionados foram: número de árvores na floresta = 500 e número de atributos selecionados para determinar a divisão em cada nó das árvores = 7.

No algoritmo K-NN o parâmetro *k* foi selecionado através da validação cruzada (10-*fold*), sendo obtido o valor de *k* = 7. Na RVFL, o número de nós de realce (neurônios da camada oculta) foi determinado pelo método da validação cruzada para evitar *overfitting*, sendo obrtido o valor de 14.

Após a seleção dos parâmetros dos modelos foi realizada a fase de treinamento e, após, os modelos obtidos foram utilizados para estimar os valores das características da qualidade utilizando os dados do grupo de teste.

A Tabela 1 apresenta os valores dos erros, calculados utilizando as equações (5) até (8), para os dados do grupo de treinamento e do grupo de teste. Para todos os modelos, os valores dos erros obtidos pelos modelos para os grupos de treinamento e de teste não diferem significativamente, indicando não haver sobre ajuste dos modelos obtidos com as técnicas de aprendizado de máquina.

Tabela 1 - Valores dos erros calculados para o grupo de treinamento e de teste

Modelo	Grupo de treinamento			Grupo de teste		
	MSE	RMSE	MAE	MSE	RMSE	MAE
RNA	0,01249	0,1117	0,08796	0,01391	0,1179	0,09431
SVR	0,01305	0,1142	0,08462	0,01355	0,1164	0,08951
RVR	0,01257	0,1119	0,08650	0,01348	0,1161	0,09060
GPR	0,01247	0,1117	0,08790	0,01391	0,1179	0,09431
RT	0,01482	0,1217	0,09685	0,01642	0,1281	0,10475
MT	0,01286	0,1134	0,08945	0,01491	0,1221	0,09760
RF	0,01243	0,1115	0,08695	0,01366	0,1168	0,09231
K-NN	0,01372	0,1171	0,09052	0,01464	0,1210	0,09397
RVFL	0,01266	0,1125	0,08790	0,01413	0,1189	0,09604

Fonte: Autoria própria (2020)

Considerando os valores dos erros apresentados na Tabela 3 pode-se considerar que todos os modelos representam adequadamente a fração de produtos não conforme do processo. Analisando os resultados apresentados na Tabela 1 verifica-se que:

a) O modelo RF apresenta os menores valores de MSE e RMSE para os dados do grupo de treinamento dentre todos os modelos;
b) O modelo SVR apresenta o menor valor de MAE para os dados do grupo de treinamento dentre todos os modelos;
c) A ordem ascendente dos valores do MSE e RMSE para os dados do grupo de treinamento é:
   RF < RNA < GPR < RVR < RVFL < MT < SVR < K-NN < RT
d) A ordem ascendente dos valores do MAE para os dados do grupo de treinamento é:
   SVR < RVR < RF < GPR < RVFL < RNA < MT < K-NN < RT
e) O modelo RVR apresenta os menores valores de MSE e RMSE para os dados do grupo de teste dentre todos os modelos;
f) O modelo SVR apresenta o menor valor de MAE para os dados do grupo de teste dentre todos os modelos;
g) A ordem ascendente dos valores do MSE e RMSE para os dados do grupo de teste é:
   RVR < SVR < RF < RNA < GPR < RVFL < K-NN < MT < RT
h) A ordem ascendente dos valores do MAE para os dados do grupo de teste é:
   SVR < RVR < RF < K-NN < RNA < GPR < RVFL < MT < RT
i) O modelo RT apresentou os maiores valores de erros, tanto para os dados do grupo de treinamento quanto de teste.

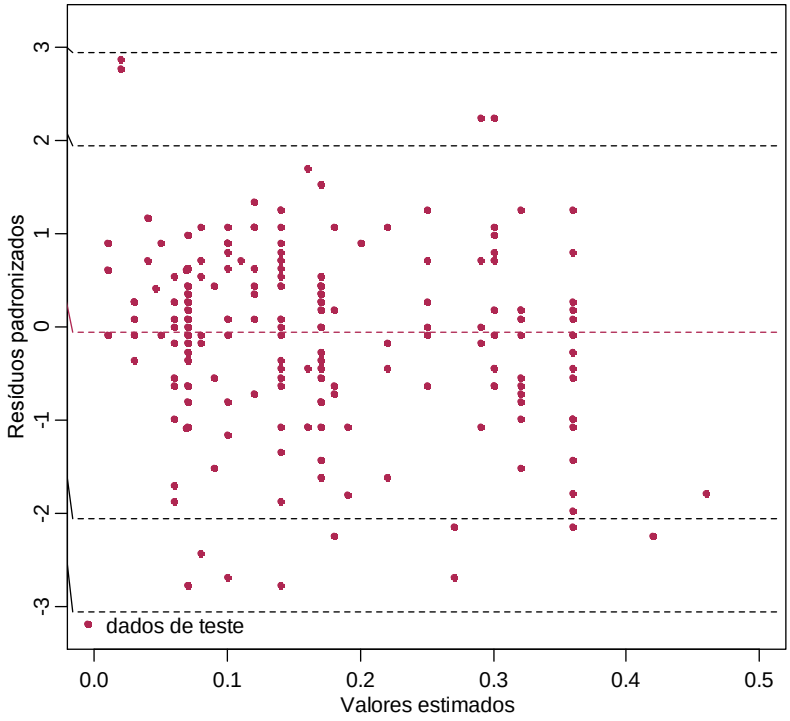
Das técnicas de aprendizado de máquina analisadas, pode-se considerar que o modelo RVR apresentou melhor desempenho para os dados do grupo de teste, com o modelo SVR em seguida. Deve-se levar em consideração, também, que no modelo SVR devem ser selecionados os parâmetros constante de regularização (*C*), função de perda insensível (ε) e o parâmetro (γ) da função kernel RBF; enquanto no modelo RVR é necessário selecionar apenas o parâmetro (γ) da função kernel RBF.

A verificação da normalidade dos resíduos dos modelos foi realizada utilizando o teste de normalidade de Shapiro-Wilk com 95% nível de confiança, sendo que pode-se considerar que os resíduos de todos os modelos seguem a distribuição normal, pois *p-value* > 0,05.

Outro teste realizado para a validação dos modelos de regressão estimados foi a estatística de Durbin-Watson, para verificar a hipótese de que os resíduos são independentes, sendo que os resultados obtidos para os modelos indica que os resíduos são independentes. Para a verificação se os resíduos possuem variância aproximadamente constante (homoscedasticidade) foi realizado o teste de Levene. Como *p-value* > 0,05 foi obtido em todos os modelos, com os dados do grupo de treinamento e de teste, não se rejeita a hipótese de homoscedasticidade dos resíduos.

Como o modelo RVR apresentou o melhor desempenho para os dados do grupo de teste, foi elaborado o gráfico dos resíduos padronizados versus os valores estimados, apresentado na Figura 1. Este gráfico pode ser utilizado para a verificação do grau de concordância entre os resíduos padronizados e os valores estimados. Se os resíduos padronizados são independentes e identicamente distribuídos (i.i.d.), em torno de 95% devem estar no intervalo [-2,2] e 99% no intervalo [-3;+3]. Verifica-se na Figura 1 que todos os pontos estão no intervalo de [-3;+3].

Figura 1 – Gráfico dos resíduos padronizados versus valores estimados pelo modelo RVR para os dados do grupo de teste



Fonte: Autoria própria (2020)

A Figura 2 apresenta o monitoramento dos resíduos padronizados utilizando os limites de controle calculados a partir dos resíduos do modelo RVR para os dados do grupo de teste. Como todos os resíduos padronizados encontram-se dentro dos limites de controle, nenhuma evidência de presença de causas especiais foi detectada e o processo pode ser considerado sob controle estatístico.

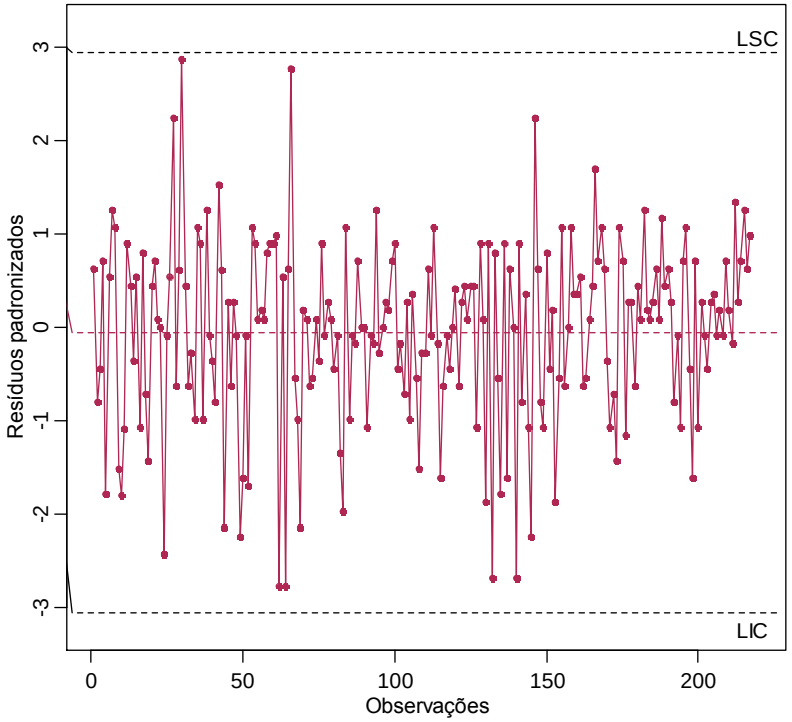
5. Conclusões

Os efeitos das variáveis independentes (fatores controláveis) sobre a variável dependente (característica da qualidade) em processos industriais podem ser analisados através da modelagem dos dados destes processos usando modelos de regressão. Para esta modelagem podem ser utilizadas técnicas estatísticas e técnicas de aprendizado de máquina.

Neste artigo foram utilizadas técnicas de aprendizado de máquina (RNA, SVR, RVR, GPR, RT, MT, RF, K-NN e RVFL) para a modelagem da fração de produtos não conformes às

especificações de uma indústria curtidora de couro em que a característica da qualidade é mensurada no intervalo [0,1]. No modelo RF foram obtidos os menores valores de MSE e RMSE para os dados do grupo de treinamento e o modelo RVR apresenta os menores valores de MSE e RMSE para os dados do grupo de teste.

Figura 2 – Cartas de controle para os resíduos padronizados obtidos com o modelo RVR para os dados do grupo de teste



Fonte: Autoria própria (2020)

Os resíduos dos modelos de regressão devem ser analisados com relação a normalidade, independência e variância. Foi realizado o teste de normalidade de Shapiro-Wilk e o teste de Levene, bem como foi calculada a estatística de Durbin-Watson. Pelos resultados obtidos verificou-se que os resíduos dos modelos seguem a distribuição normal, não se rejeita a hipótese de homoscedasticidade dos resíduos e os resíduos são independentes.

Como o modelo RVR apresentou o melhor desempenho para os dados do grupo de teste, foi elaborado o gráfico dos resíduos padronizados versus os valores estimados, onde verificou-se que os resíduos padronizados são i.i.d., e foi realizado o monitoramento dos resíduos padronizados, sendo que todos os resíduos padronizados encontram-se dentro dos limites de controle.

Através dos resultados pode-se considerar que os modelos baseados nas técnicas de aprendizado de máquina representam adequadamente a fração de produtos não conforme do processo produtivo.

Referências

BENEDETTI, Jacqueline K. On the nonparametric estimation of regression functions. **Journal of the Royal Statistical Society**. Series B (Methodological), v. 39, n. 2, p. 248-253, 1977.

BISHOP, Christopher M. **Pattern recognition and machine learning**. New York: Springer, 2006.

BRAGA, Antônio de P.; CARVALHO, André P. de L. F.; LUDERMIR, Teresa B. **Redes neurais artificiais**: teoria e aplicações. 2. ed. Rio de Janeiro: LTC, 2012.

BREIMAN, Leo. Random forests. **Machine Learning**, v. 45, p. 5-32, 2001.

BREIMAN, L.; FRIEDMAN, J. H.; OLSHEN, R. A.; Stone, C.J. **Classication and regression trees**, Belmont, California: Wadsworth, 1984.

ELITH, J.; LEATHWICK, J. R.; HASTIE, T. A working guide to boosted regression trees. **Journal of Animal Ecology**, v. 77, n. 4, p. 802-813, 2008.

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

FERNÁNDEZ-DELGADO, M.; SIRSAT, M.S.; CERNADAS, E.; ALAWADI, S.; Barro, S.; FEBRERO-BANDE, M. An extensive experimental survey of regression methods, **Neural Networks**, v. 111, p. 11-34, 2019.

HAYKIN, Simon. **Neural networks and learning machines**. 3 ed. New York: Prentice Hall, 2009.

IZBICKI, Rafael; SANTOS, Tiago M. **Machine learning sob a ótica estatística**, 2019. Disponível em: <http://www.rizbicki.ufscar.br/sml.pdf >. Acesso em: 20 abril 2019.

KECMAN, Vojislav. **Learning and soft computing**: support vector machines, neural networks and fuzzy logic models. London: MIT Press, 2001.

LIAW, Andy; WIENER, Matthew. Classification and regression by random forest. **R News**, v. 2, p. 18-22, 2002.

MACKAY, David J. C. Introduction to gaussian processes. In: Bishop, Christopher M. (Org.). **Neural Networks and Machine Learning**. Springer-Verlag, 1998.

MITCHELL, Tom M. **Machine learning**. New York: McGraw-Hill, 1997.

MONTGOMERY, Douglas C.; RUNGER, George C. **Estatística aplicada e probabilidade para engenheiros**. Rio de Janeiro: LTC, 2003.

MONTGOMERY, Douglas C. **Introdução ao controle estatístico da qualidade**. Rio de Janeiro: LTC, 2004.

PAO, Y.-H.; PARK, G.-H.; SOBAJIC, D. J. Learning and generalization characteristics of the random vector functional-link net. **Neurocomputing**, v. 6, n. 2, p. 163-180, 1984.

PROBST, Philipp; BOULESTEIX, Anne-Laure. To tune or not to tune the number of trees in random forest. **Journal of Machine Learning Resource**, v. 18, n. 181, p. 1-18, 2018.

QUINLAN, J. R. Learning with continuous classes. **Proceedings of the 5th Australian Joint Conference on Artificial Intelligence**, p. 343-348, 1992.

QUINLAN, J. R. Combining instance-based and model-based learning. **Proceedings of the Tenth International Conference on Machine Learning**, p. 236-243, 1993.

QIU, X.; SUGANTHAN, P. N.; AMARATUNGA, G. A. J. Ensemble incremental learning Random Vector Functional Link network for short-term electric load forecasting. **Knowledge-Based Systems**, v. 145, p. 182-196, 2018.

RASMUSSEN, Carl E.; WILLIAMS, Christopher K. I. **Gaussian processes for machine learning**. MIT Press, 2006.

SAFAVIAN, S. Rasoul; LANDGREBE, David. A survey of decision tree classifier methodology. **IEEE Transactions on Systems, Man, and Cybernetics**, v. 21, n. 3, p. 660-674, 1991.

SCHÖLKOPF, Bernhard; SMOLA, Alex J. A tutorial on support vector regression. **Statistics and Computing**, Netherlands, v. 14, n. 3, p. 199-222, 2004.

SCHULZ, Eric; SPEEKENBRINK, Maarten; KRAUSE, Andreas. A tutorial on gaussian process regression: modelling, exploring, and exploiting functions. **Journal of Mathematical Psychology**, vol. 85, p. 1–16, 2018.

TIPPING, Michael E. The relevance vector machine. In: SOLLA, S. A.; LEEN, T. K.; MÜLLER, K. R. (Org**.), Advances in neural information processing systems**, v. 12, p. 652-658, MIT Press, 2000.

TIPPING, Michael E. Sparse Bayesian learning and the relevance vector machine. **Journal of Machine Learning Research**, v. 1, p. 211-244, 2001.

VAPNIK, Vladimir N. **Statistical learning theory**. New York: John Wiley & Sons, 1998.

WERKEMA, Maria C. C.; AGUIAR, Sílvio. **Análise de regressão**: como entender o relacionamento entre as variáveis de um processo. Belo Horizonte: Fundação Christiano Ottoni, 1996.

ZHANG, C.; HE, Y.; YUAN, L.; XIANG, S.; WANG, J. Prognostics of lithium-ion batteries based on wavelet denoising and DE-RVM. **Computational Intelligence and Neuroscience**, p. 1-8, 2015.