MÉTODO HÍBRIDO DE CLASSIFICAÇÃO BINÁRIA BASEADO EM ENSEMBLE LOGISTIC GMDH-TYPE NEURAL NETWORKS

João B.G. Brito
Universidade Federal do Rio Grande do Sul – Departamento de Engenharia de Produção
Av. Osvaldo Aranha, 99 – 5° andar, Porto Alegre, RS, Brasil
joao.brito@ufrgs.br

Guilherme B. Bucco
Universidade Federal do Rio Grande do Sul – Escola de Administração
Washington Luiz, 855 – Porto Alegre, RS, Brasil
gbbucco@gmail.com

Juliana A. Fontes
Universidade Federal do Rio Grande do Sul – Departamento de Engenharia de Produção
Av. Osvaldo Aranha, 99 – 5° andar, Porto Alegre, RS, Brasil
juliana.abreu.fontes@gmail.com

Michel J. Anzanello
Universidade Federal do Rio Grande do Sul – Departamento de Engenharia de Produção
Av. Osvaldo Aranha, 99 – 5° andar, Porto Alegre, RS, Brasil
michel.anzanello@gmail.com

RESUMO

Processos industriais e químicos geram grandes volumes de variáveis, que podem apresentar níveis de ruído e correlações que dificultam o seu monitoramento e controle. Este estudo propõe um novo método para identificação do subconjunto de variáveis com maior poder de classificação de amostras industriais. Para tanto, a decomposição de valores singulares (SVD) é integrada com a técnica *Ensemble Logistic* GMDH-type Neural Networks (EL-GMDH-NN). O SVD é utilizado para calcular o índice de importância de cada variável. Um processo iterativo cria modelos adicionando variáveis na ordem decrescente do índice de importância. As variáveis que aumentam a acurácia de classificação são selecionadas. A abordagem produziu classificações 100% acuradas com 2.72% do banco de dados OXY. Na base de dados referente ao ponto de fulgor do biodiesel/diesel, 98.8% de acurácia com 0.33% das 1.738 variáveis. Os resultados foram superiores quando comparados com Regressão Logística Stepwise, *Logistic* GMDH e Anzanello *et al*. (2009).

Palavras-chave: Ranking não supervisionado. Stacking Ensemble. Group Method Data Handling.

Tópicos: (a) Redução de dimensionalidade dos dados com seleção de variáveis; (b) propomos um índice de importância de variáveis com base na decomposição de valor singular; e, (c) utilização de heurística com Group Method Data Handling.

ABSTRACT

Industrial and chemical processes generate large volumes of features, which can present noise levels and correlations that make their monitoring and control difficult. This study proposes a new method to find the subset of features with greater power of industrial sample classification. For this purpose, the decomposition of singular values (SVD) is integrated with the Ensemble Logistic GMDH-type Neural Networks (EL-GMDH-NN) technique. The SVD is used to calculate the importance index of each feature. An iterative process creates models by adding features in descending order of importance index. features that increase the accuracy of classification are selected. The approach produced 100% accurate rankings with 2.72% of the OXY database. In the biodiesel/diesel flash point database, 98.8% accuracy with 0.33% of 1,738 features. The results were superior when compared with Stepwise Logistic Regression, Logistic GMDH and Anzanello *et al*. (2009).

Galoá { Este trabalho foi publicado utilizando o **Galoá** proceedings

Keywords: Unsupervised Ranking. Stacking Ensemble. Group Method Data Handling.

Highlights: (a) Data dimensionality reduction with variable selection; (b) we propose an index of importance of variables based on the decomposition of singular value; e, (c) use of heuristics with Group Method Data Handling.

1 Introdução

O processo de seleção de variáveis não é trivial, pois raramente as variáveis são totalmente independentes, sendo comum possuírem redundância e baixo nível de informação (HUNTER, 2000). De acordo com Anzanello *et al*. (2009), processos industriais e químicos geram um elevado número de variáveis, muitas delas correlacionadas. Bancos de dados podem ter variáveis irrelevantes com potencial para interferir nas variáveis reais, introduzindo heterogeneidade nos dados e gerando dependência (KHAIRE; DHANALAKSHMI, 2019). No entanto, a verificação de todos os subconjuntos possíveis de varáveis pode ser extremamente onerosa (precisamente $2^n - 1$, onde n é o número de variáveis) e rapidamente convergir para um problema NP-difícil, em razão do grande número de subconjuntos possíveis (JENSEN; SHEN, 2008).

Modelos para seleção de variáveis podem implementar estratégias que mesclam diferentes técnicas. Li *et al*. (2014) combinam estatística com aprendizagem de máquina para definir as variáveis mais importantes em processos financeiros do gerenciamento de risco corporativo. Anzanello *et al*. (2013) propõem um método de seleção de variáveis com uma abordagem heurística integrada, no qual aplica-se inicialmente a análise de componentes principais nos dados originais para criar um índice de importância das variáveis. Em seguida, modelos *k-nearest neighbor* são criados com a *backward elimination* das variáveis. Por fim, é selecionado o subconjunto de variáveis que compõem o modelo mais acurado. Comparado ao PCA, o SVD tem a vantagem de usar matrizes retangulares sem a necessidade de pré-processamento para transformação de dados em matrizes de covariância ou matrizes de correlação (GOLUB; KAHAN, 1965). Ademais, Skillicorn (2007) destaca que o SVD analisa os atributos de forma conjunta, o que resulta em mais sensibilidade na identificação do ruído sobre todo o conjunto de dados. Da mesma forma, no âmbito da aprendizagem de máquina, métodos como Redes Neurais Artificiais (ANN) são uma alternativa para o KNN, em geral apresentando desempenho superior (MOOSAVIAN *et al*., 2013).

O presente artigo implementa um novo método, denominado SVD-GMDH, que seleciona variáveis para classificação binária de amostras oriundas de processos industriais. A proposta utiliza o SVD para calcular o índice de importância de cada variável e o integra ao método *Ensemble Logistic GMDH-type Neural Networks* (EL-GMDH-NN). O EL-GMDH-NN tem como principal vantagem agregar o desempenho de múltiplos algoritmos de aprendizagem, combinando-os em um único algoritmo de aprendizagem, denominado *meta learning*, como a rede neural logística *Logistic* GMDH-NN (KONDO *et al.*, 1999). A *Logistic* GMDH-NN otimiza o número de neurônios, camadas e variáveis de entrada mais relevantes (KONDO, 1998). Essa característica confere desempenho superior frente aos métodos tradicionais de redes neurais (AHMDI *et al*, 2019). Ademais, quando o GMDH-NN é utilizado como *meta learning*, seu desempenho tende a ser ainda mais destacável (XIAO *et al.*, 2014).

Para analisar o desempenho da proposta SVD-GMDH foram aplicados experimentos sobre diversos conjuntos de dados caracterizados por uma classe binária de qualidade: conforme ou não conforme. Parte dos bancos se referem a processos de produção industrial, incluindo produção de nylon, emulsão na indústria de papel, polimerização em um processo de látex, produção de óxido de titânio e produção de antibióticos. Nesse coletivo de dados, a identificação do subconjunto de variáveis que mais contribui para correta classificação, entre um produto conforme ou não conforme, também fornece informações que possibilitam a revisão do processo de aferição. Essa análise pode levar a redução de custos no processo de apuração da qualidade, caso variáveis possam

Galoá { Este trabalho foi publicado utilizando o **Galoá** proceedings

ser descartadas e não aferidas. O segundo conjunto de dados analisado está relacionado a propriedades físico-químicas do Biodiesel e Diesel brasileiro. Nessa conjuntura, um algoritmo que classifica com eficiência amostras entre conformes ou não conformes oferece confiabilidade ao processo de classificação. Biodiesel e Diesel de má qualidade refletem diretamente em desperdício de combustível, desgaste acelerado do motor e custos com manutenção do veículo (Ferrão *et al.*, 2011). Além disso, tanto para o conjunto do Biodiesel e Diesel quanto para os processos industriais, a classificação incorreta de conformidade tende a gerar leso material e financeiro ao cliente (BRASIL, 1990) e pode inclusive impactar na imagem de uma marca (PASQUALINI, 2019).

2 Método proposto – SVD-GMDH

O desenvolvimento da proposta é organizado em três fases: (i) divisão do banco de dados, (ii) criação do índice de importância, e (iii) seleção das variáveis, conforme figura 2.1.

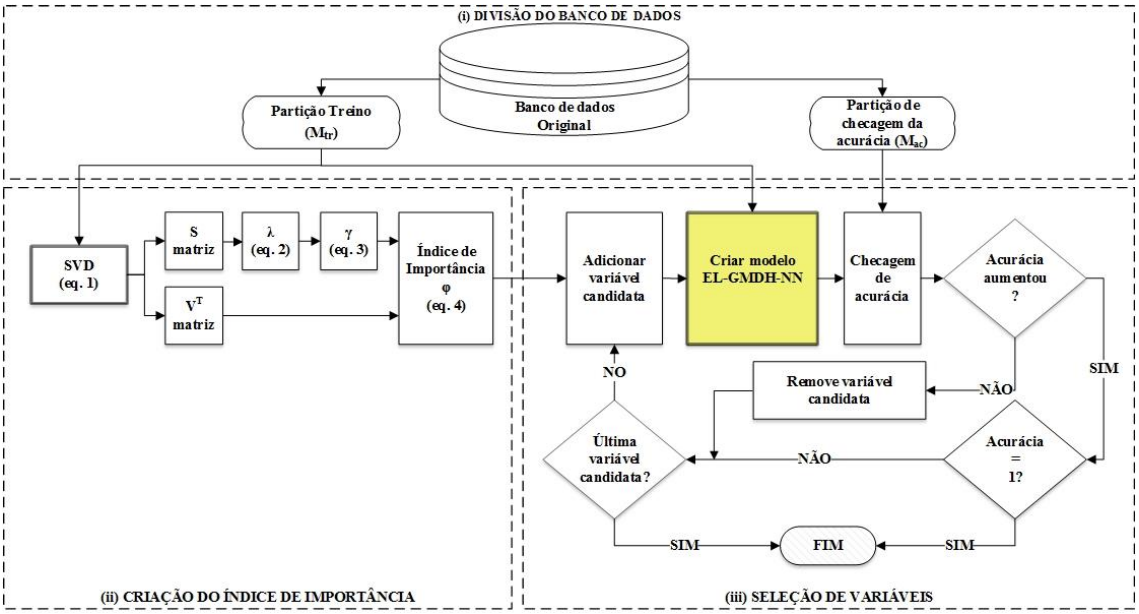


Figura 2.1 - Fluxograma do modelo SVD-GMDH

2.1 Divisão do banco de dados

Considera-se uma matriz com M observações, descritas por uma classe binária "0", que indica observação com qualidade não conforme e "1", que representa observação com qualidade conforme, contendo N variáveis com valores reais. Tal matriz é aleatoriamente dividida em duas partições: M_{tr} para treino e M_{ac} de teste, para teste, com $M_{\mathrm{tr}} + M_{\mathrm{ac}} = M$.

2.2 Criação do índice de importância

Skillicorn (2007) define o SVD como um método de decomposição matricial que transforma uma matriz $\mathrm{A}_{|\mathrm{mn}|} \in \mathbb{R}_{|\mathrm{mn}|}$, com M linhas e N colunas, e rank r, de forma a expor a quantidade de variância de um conjunto de dados utilizando variáveis latentes. Para o autor, a interpretação mais intuitiva é a geométrica: os dados da matriz $\mathrm{A}_{|\mathrm{mn}|}$ são transformados de tal forma que a variância é maximizada em um eixo; o método então posiciona os dados em um novo eixo que é, em relação ao primeiro, ortogonal e com variância maximizada; o método continua adicionando novos eixos ortogonais com variância maximizada a todos os anteriores, até r da matriz de dados. A fatoração SVD da matriz $\mathrm{A}_{|\mathrm{mn}|}$ é representada pelo produto de três matrizes $\mathrm{U}_{|\mathrm{rr}|}$, $\mathrm{S}_{|\mathrm{rr}|}$ e $\mathrm{V}^{\mathrm{T}}_{|\mathrm{rn}|}$ (equação 2.1). As matrizes $\mathrm{U}_{|\mathrm{rr}|}$ e $\mathrm{V}^{\mathrm{T}}_{|\mathrm{rn}|}$ são ortogonais esquerda e direita, respectivamente, e a matriz $\Delta_{|\mathrm{rr}|} \in \mathbb{R}_{+}$ é diagonal.

Galoá { Este trabalho foi publicado utilizando o Galoá proceedings

$$A_{|mn|} = U_{|rr|}\Delta_{|rr|}V^{T}_{|rn|} \quad (2.1)$$

Onde $A_{|mn|}$ é a matriz original dos dados, $U_{|rr|}$ é a matriz com as coordenadas dos objetos no espaço estendido pelos novos eixos, $\Delta_{|rr|}$ é a matriz de valores singulares com fator de escala que indica a importância relativa de cada novo eixo e, $V^{T}_{|rn|,}$ é a matriz com os novos eixos. A partir do SVD calcula-se a quantidade de variância explicada, γ. Para calcular λ, é necessário calcular os autovalores λ, de acordo com equação 2.2.

$$\lambda_j = \mu_j^2 (j = 1, \dots, r) \quad (2.2)$$

onde, r é o número de valores singulares, j é o índice de valores singulares e autovalores correspondentes, λ_j é o *j*-ésimo autovalor e μ_j é o *j*-ésimo valor singular.

O *j*-ésimo valor singular μ_j é determinado de tal forma que a variância entre os valores singulares é maximizada. A quantidade de variação explicada por cada dimensão singular é representada pelo coeficiente γ_j, como na equação 2.3.

$$\gamma_j = \frac{\lambda_j}{\sum_{j=1}^{r} \lambda_j} \quad (2.3)$$

Então, o índice de importância é calculado, conforme equação 2.4.

$$\varphi_a = \sum_{j=1}^{r} \gamma_j v_{aj}^2, a = 1, \dots, N \quad (2.4)$$

Por fim, o índice de importância é aplicado para ordenar as variáveis de forma decrescente (maiores valores de índice sugerem variáveis mais relevantes para o processo de classificação).

2.3 Seleção de variáveis

A fase de seleção de variáveis ocorre em um processo iterativo, no qual cada variável candidata é adicionada ao modelo *Ensemble Logistic* GMDH-NN de acordo com a abordagem de inclusão *forward*. A acurácia do modelo é verificada com a partição M_{ac}. Se houver aumento na acurácia, a variável candidata é mantida no subconjunto de variáveis selecionadas. Quando todas as variáveis forem testadas ou o subconjunto em análise atingir acurácia de 100%, o processo é encerrado. Como resultado, é obtido o subconjunto de variáveis recomendado para predição. O modelo EL-GMDH-NN utiliza a metodologia ensemble e é descrito, conforme figura 2.2, em cinco passos: (i) Divisão da partição de treino M_{tr} em sub-partições M_{sub-tr} e M_{sub-ts}; (ii) *Learners*: algoritmos de aprendizagem de máquina treinados com o subconjunto de variáveis selecionadas e a nova variável candidata, adicionada conforme ordem do índice de importância; (iii) *Predictions*: executar as predições utilizando a sub-partição M_{sub-ts}; (iv) Ensemble Logistic GMDH-NN: treinamento do modelo com a predição de cada *learner* e a binária classe da partição M_{sub-ts}.

Galoá { Este trabalho foi publicado utilizando o **Galoá** proceedings

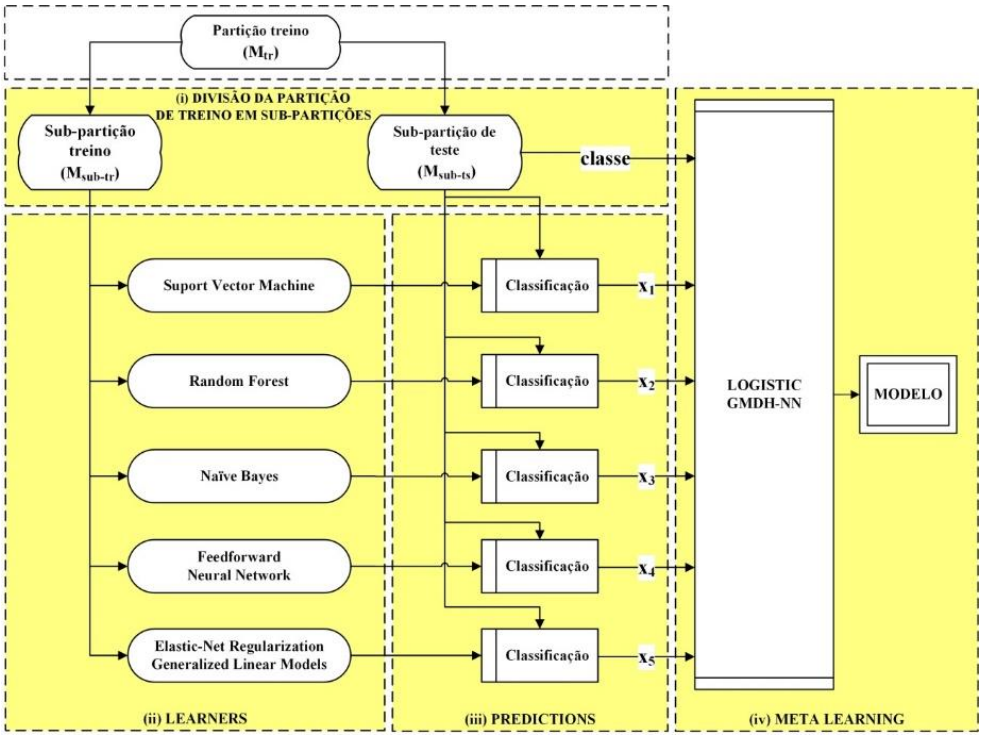


Figura 2.2 - Criação do modelo EL-GMDH-NN

2.3.1 *Learners* e *predictions*

A criação do modelo Ensemble Logistic GMDH-NN inicia com divisão aleatória da partição M_{tr} nas sub-partições $M_{\text{sub}-\text{tr}}$ e $M_{\text{sub}-\text{ts}}$, na razão de 70-30%. Então a partição $M_{\text{sub}-\text{tr}}$, que contém somente as variáveis eleitas e a nova candidata, é utilizada para modelos *learners*: *Suport Vector Machine* (SVM), *Random Forest* (RF), *Naïve Bayes* (NB), *Regularization Paths for Generalized Linear Models Coordinate Descent* (GLMNET) e *Single-hidden-layer neural network* (NNET) (RIPLEY, 2007). Os *learners* treinados são utilizados para predição da sub-partição $M_{\text{sub}-\text{ts}}$. O *meta learning* Logistic GMDH-NN é criado utilizando como classe de resposta à classe da sub-partição M_{ts} e como preditores o resultado da predição de cada um dos *learners*. A predição de cada um dos *learners*, SVM, RF, NB, GLMNET e NNET, corresponde a uma variável independente do *meta learning*, representada, respectivamente, por x_1, x_2, x_3, x_4 e x_5.

2.3.2 Meta learning - logistic GMDH-NN

Logistic GMDH-NN é um modelo que integra o método de redes neurais artificiais Feedforward com a heurística de auto-organização GMDH (IVAKHNENKO, 1970). Na primeira camada do modelo de rede neural, cada variável preditiva é representada por um neurônio, então pares de neurônios são conectados via polinômio quadrático Kolmogorov–Gabor, resultando em $n^n - 1$ combinações, sendo n o número de variáveis de entrada (NARIMAN- ZADEH *et al*., 2002). Em seguida, um critério externo (geralmente erro médio quadrado (MSE)), também conhecido como Criterion for Regularity (CR) é calculado para cada conjunto de testes, de acordo com a equação 2.5 (FERNÁNDEZ; LOZANO, 2010). Conforme os autores, somente os neurônios com o menor valor de CR são mantidos. Em seguida, uma nova camada é criada usando as saídas selecionadas e o processo começa novamente formando pares de entradas. Essas etapas são repetidas para gerar novas camadas até que o critério de erro pare de diminuir. Quando isso ocorre, a saída obtida será a que atingiu o menor valor de critério de erro na camada anterior.

Galoá { Este trabalho foi publicado utilizando o **Galoá** proceedings

$$CR = \frac{1}{T} \sum_{n=1}^{T} (\hat{y}_n - y_n)^2 \quad (2.5)$$

Onde T é o número de vetores do subconjunto de teste; $\hat{y}_n$ é o valor predito para a *n*-ésima variável; e y_n é o valor atual da *n*-ésima variável. Os parâmetros da rede são estimados utilizando séries de Volterra (VOLTERRA, 1930; MASOUMNEZHAD *et al.*, 2016). A equação 2.6 demonstra a conexão entre as entradas e saídas do modelo.

$$\hat{y} = a_0 + \sum_{i=1}^{m} b_i x_i + \sum_{i=1}^{m}\sum_{j=1}^{m} c_{ij} x_i x_j + \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{m} d_{ijk} x_i x_j x_k + \cdots \quad (2.6)$$

Onde $\hat{y}$ é a predição; $\hat{y} \in \{0, 1\}$; m o número de entradas $a, b, c, d, e, \cdots$ são os coeficientes da função; x é o vetor de entrada variáveis e $i, j, k, \cdots$ são os índices das variáveis. A saída $\hat{y}$ é arredondada para zero quando valor for menor que 0,5, caso contrário é assumido o valor 1.

A acurácia do modelo EL-GMDH-NN indica se o processo deve continuar, adicionar uma nova variável ou ser finalizado. Esse passo utiliza a partição M_{ac} para extrair a predição dos *learners* treinados e submeter às predições ao *meta learning* EL-GMDH-NN. O processo é finalizado quando a acurácia atinge o valor 1 ou quando não há mais variáveis candidatas (o que acontecer antes).

3 Experimentos, resultados e discussões

3.1 Amostras

O método proposto SVD-GMDH foi aplicado sobre nove bancos de dados padronizados. Cinco foram obtidas por Gauchi e Chagnon (2001) e contêm dados relacionados a aferições de processos industriais, como a produção de nylon (ADPN), emulsão na indústria de papel (GRANU), polimerização em um processo de látex (LATEX), produção de óxido de titânio (OXY) e produção de antibióticos (SPIRA). As variáveis preditoras a serem selecionadas descrevem temperaturas, pressões e concentração de reagentes químicos dos processos analisados. A variável de resposta (classe) indica uma característica do produto, como viscosidade ou pureza, e é definida em dois níveis (conforme ou não conforme), de acordo com especificações de Gauchi e Chagnon (2001). Os outros quatro bancos analisados foram fornecidos por Ferrão *et al.* (2011) e são relacionados a propriedades físico-químicas do biodiesel e diesel brasileiro, como temperatura de transição do estado líquido para o gasoso (BOILING POINT), velocidade de combustão e compressão para ignição (CETANE NUMBER), temperatura de ignição instantânea em contato com fogo (FLASH POINT) e viscosidade do combustível (VICOSITY). As classes foram definidas com base na mediana da variável resposta. Valores acima da mediana foram considerados conformes e valores abaixo ou iguais à mediana foram considerados como não conformes em relação à qualidade. A tabela 3.1 relaciona o nome de cada banco analisado, natureza da aplicação, quantidade de variáveis, quantidade de observações, quantidade de observações classificadas como conforme e como não conforme e número de observações em cada uma das partições de treino-checagem de acurácia.

Tabela 3.1 - Bancos de dados de processos industriais e propriedades físico-químicas do Biodiesel e Diesel brasileiro

Banco de dados	Natureza da aplicação	Número total de variáveis	Número total de observações	Conformidade		Número de observações					
						Partição de treino			Partição de teste		
				Sim	Não	60%	75%	90%	40%	25%	10%

Galoá { Este trabalho foi publicado utilizando o Galoá proceedings

ADPN	Produção de nylon	100	71	51	20	43	53	64	28	18	7
GRANU	Emulsão na indústria de papel	78	29	15	14	17	22	26	12	7	3
LATEX	Polimerização em um processo látex	117	262	184	78	157	197	236	105	66	26
OXY	Produção de óxido de titânio	95	25	18	7	15	19	23	10	6	3
SPIRA	Produção de antibióticos	96	145	95	50	87	109	131	58	36	15
BOILING POINT (C°)	Temperatura de transição do estado líquido para o gasoso do Biodiesel e Diesel e brasileiro	401	113	60	53	68	85	102	45	28	11
CETANE NUMBER	Indicador de velocidade de combustão e compressão para ignição do Biodiesel e Diesel e brasileiro	401	113	59	54	68	85	102	45	28	11
FLASH POINT (C°)	Temperatura de ignição instantânea em contato com fogo do Biodiesel e Diesel e brasileiro	1738	85	44	41	51	64	77	34	21	9
VICOSITY	Viscosidade do combustível do Biodiesel e Diesel e brasileiro.	401	116	59	57	70	87	104	46	29	12

3.2 Experimentos

Os resultados foram comparados com o estudo de Anzanello *et al.* (2009), o qual implementou um novo índice de importância das variáveis baseado em Mínimos Quadrados Parciais (PLS). No estudo dos autores, a proposta foi comparada com outras abordagens e combinada com técnicas de aprendizagem de máquina, tais como: *k-Nearest Neighbor* (KNN), *Probabilistic Neural Network* (PNN) e *Suport Vector Machine* com regressão PLS. Ademais, para comparar os resultados do método, foram utilizadas outras duas técnicas: *Logistic Stepwise Regression* (*bidirectional elimination*) e *Logistic* GMDH-NN (KONDO *et al.*, 1999). Os experimentos foram conduzidos com porções de treino-checagem de acurácia gerados aleatoriamente de forma estratificada e proporcional por classe nos conjuntos 60-40%, 75-25% e 90-10%. No estudo de Anzanello *et al.* (2009) não estão incluídos experimentos com partições 60%-40% e 75-25%, dessa forma há lacunas de comparação para essas partições. Foram executadas de 200 a 10.000 simulações independentes, no qual o critério de parada foi a convergência da média de acurácia das últimas dez iterações. Para implementar a proposta foi utilizada a linguagem de programação R (versão 3.6.3) (R CORE TEAM, 2019), com o pacote GMDH2 (versão 1.4) (DAG *et al.*, 2019).

3.3 Resultados e discussões

Foram apurados como resultados o percentual médio de variáveis retidas (% variáveis retidas), percentual médio de acurácia (% acurácia) e percentual médio do coeficiente de variação da acurácia (% CF) em percentual. A tabela 3.2 demonstra o desempenho médio do SVD-GMDH para cada banco de dados em cada partição de treino-checagem de acurácia.

Tabela 3.2 - Desempenho médio do SVD-GMDH

Banco de dados	Porções do banco de dados % treino–% checagem da acurácia		
	60-40%	75-25%	90-10%

Galoá { Este trabalho foi publicado utilizando o Galoá proceedings

	% Acurácia (% CF)	% Variáveis retidas	% Acurácia (%CF)	% Variáveis retidas	% Acurácia (%CF)	% Variáveis retidas
ADPN	91,95 (5,06)	6,45	93,25 (4,87)	5,61	96,64 (5,75)	3,81
GRANU	88,3 (5,1)	2,34	90,28 (5,86)	1,99	94,2 (5,94)	1,37
LATEX	83,98 (5,34)	2,08	85,34 (6,28)	1,88	91,97 (6,95)	1,41
OXY	99,63 (0,98)	0,55	99,72 (1,09)	0,43	99,8 (1,47)	0,33
SPIRA	90,42 (4,24)	2,31	93,06 (4,47)	2,05	97,62 (4,79)	1,59
BOILING POINT	89,7 (6,3)	6,46	93,35 (3,82)	6,05	100 (0)	4,1
CETANE NUMBER	86,3 (3,58)	8,39	87,47 (4,48)	7,5	90,93 (5,86)	5,45
FLASH POINT	100 (0)	3,99	100 (0)	3,28	100 (0)	2,72
VICOSITY	83,24 (4,74)	8,54	86,02 (5,27)	7,39	91,98 (8,32)	5,62

O melhor resultado foi obtido na amostra FLASH POINT, para o qual SVD-GMDH gerou 100% de acurácia, com percentual médio de variáveis variando de 3,99%, para porção 60-40%, à 2,72%, na porção 90-10%. Na amostra BOILING POINT, a porção 90-10% também atingiu 100% de acurácia utilizando somente 4,1% das variáveis. Também destaca-se o resultado obtido na amostra OXY, que mesmo possuindo somente 25 observações, alcançou 100% de acurácia na porção 90-10%. Em geral, a acurácia foi superior a 90% em 63% dos experimentos, sendo que o menor resultado foi de 83,24% de acurácia para a amostra VISCOSITY (porção 60-40%). Da mesma forma, os resultados mostraram-se positivos em relação à variabilidade da acurácia, com 85% dos apontamentos apresentando valores inferiores a 6%. Na contramão, a variabilidade mais expressiva foi de 8,32% para porção 90-10% da amostra VICOSITY. Com relação ao percentual de variáveis retidas, 74% dos resultados foram formados por subconjuntos com menos de 6% das variáveis. O resultado mais impactante foi com a amostra de OXY, no qual, em média, somente 0,33% das variáveis foram necessárias para formar um subconjunto de preditores que gerasse, em média, 99,8% de acurácia. Outro resultado destacável ocorreu com o banco de dados FLASH POINT, que possui a mais alta dimensionalidade das bases analisadas (1738 variáveis). O método utilizou, em média, somente 3,9%, para porção 60-40%, 3,28%, na porção 75-25% e 2,72% das variáveis, com a porção 90-10%, para compor subconjuntos de preditores que produzissem classificações perfeitas. A tabela 3.3 traz o desempenho do SVD-GMDH comparado a outras técnicas.

Tabela 3.3 - Comparação de desempenho do SVD-GMDH

Banco de dados	% Treino	Anzanello *et al.* (2009)		Regressão Logística *Stepwise*		Logistic GMDH-NN		SVD-GMDH	
		% Acurácia	% variáveis retidas	% Acurácia (% CF)	% variáveis retidas	% Acurácia (% CF)	% variáveis retidas	% Acurácia (% CF)	% variáveis retidas
ADPN	60			69,82 (13,43)	6,76	79,54 (9,29)	8,06	91,95 (5,06)	6,45
	75			73,14 (14,46)	8,11	82,23 (11,14)	7,93	93,25 (4,87)	5,61
	90	87,00	8,00	74,10 (19,50)	8,62	80,82 (17,53)	8,46	**96,64 (5,75)**	**3,81**
GRANU	60			62,12 (7,00)	7,11	73,13 (10,66)	2,31	88,3 (5,1)	2,34
	75			55,55 (11,85)	5,84	74,42 (11,84)	2,23	90,28 (5,86)	1,99
	90	87,00	7,70	66,66 (0)	7,58	71,29 (22,97)	2,15	**94,2 (5,94)**	**1,37**
LATEX	60			83,83 (5,13)	18,62	70,89 (10,03)	2,17	83,98 (5,34)	2,08
	75			85,68 (4,46)	21,80	71,57 (10,93)	2,03	85,34 (6,28)	1,88

Galoá { Este trabalho foi publicado utilizando o **Galoá** proceedings

	90	83,00	18,5	81,40 (4,59)	20,09	70,89 (17,14)	2	**91,97 (6,95)**	**1,41**
OXY	60			66,66 (24,74)	2,74	98,76 (1,57)	0,63	99,63 (0,98)	0,55
	75			76,00 (7,20)	1,75	97,64 (2,37)	0,78	99,72 (1,09)	0,43
	90	73,00	6,30	83,33 (23,09)	2,11	**100 (0)**	**0,62**	99,8 (1,47)	0,33
SPIRA	60			71,92 (8,50)	16,73	70,09 (7,61)	2,36	90,42 (4,24)	2,31
	75			58,80 (15,96)	17,96	77,93 (8,66)	2,72	93,06 (4,47)	2,05
	90	90,00	4,20	61,60 (21,70)	26,42	82,29 (10,38)	2,03	**97,62 (4,79)**	**1,59**
BOILING POINT	60			70,48 (7,19)	3,45	70,24 (11,12)	10,62	89,7 (6,3)	6,46
	75			82,75 (0)	4,12	68,27 (9,5)	8,38	93,35 (8,82)	6,05
	90			87,74 (4,81)	4,25	83,33 (21,91)	3,85	**100 (0)**	**4,1**
CETANE NUMBER	60			64,49 (4,24)	3,99	78,1 (6,33)	5,98	86,3 (3,58)	8,39
	75			65,51 (0)	3,62	76,68 (0,99)	8,07	87,47 (4,48)	7,5
	90			70,83 (17,85)	5,49	81,7 (4,66)	7,29	**90,93 (5,86)**	**5,45**
FLASH POINT	60			55,88 (31,58)	100	100 (0)	9,61	100 (0)	3,99
	75			86,36 (10,52)	100	85,71 (0)	8,98	100 (0)	3,28
	90			77,77 (14,28)	100	82,46 (20,74)	10,03	**100 (0)**	**2,72**
VISCOSITY	60			80,85 (5,41)	100	74,34 (4,77)	8,88	83,24 (4,74)	8,54
	75			55,17 (6,25)	100	77,3 (5,42)	7,85	86,02 (5,27)	7,39
	90			58,33 (14,29)	100	74,29 (9,22)	7,44	**91,98 (8,32)**	**5,62**

Em todas as comparações o desempenho do SVD-GMDH mostrou-se superior com relação a acurácia, com exceção da base OXY, no qual a porção 90-10% obteve 99,8%, enquanto que no método Logistic GMDH-NN as classificações foram perfeitas. No entanto, também é possível constatar que foram necessárias, aproximadamente, o dobro de variáveis, com 0,62% contra 0,33% do SVD-GMDH. Esse resultado, em especial, indica que o índice de importância de variáveis calculado a partir do SVD desempenha um papel importante na orientação da composição dos subconjuntos. Sobre essa ótica, e ponderando que a diferença de acurácia não é extremamente pequena (0,02%), pode-se considerar que o SVD-GMDH também se mostra superior nesse resultado.

No conjunto de dados relacionados aos processos industriais (ADPN, GRANU, LATEX, OXY e SPIRA), verificou-se que o segundo método com melhor desempenho foi Anzanello *et al.* (2009). Todavia, enfatiza-se que o SVD-GMDH demonstrou desempenho superior, mesmo ao comparar a partição 60-40% (com acurácia de 90,86%) com 90-10% de Anzanello *et al.* (2009), que registrou 84% de acurácia utilizando uma porção maior da amostra para treinamento. Dentre todos os métodos, para todas as bases de dados, o pior desempenho foi obtido pelo método Regressão Logística *Stepwise*.

O processo de seleção de variáveis tem como principal objetivo a definição de um subconjunto ótimo ou subótimo de preditores, como relata Guyon e Elisseeff (2003). Porém, um grande benefício que pode ser obtido, principalmente em processos industriais, é a identificação

Galoá { Este trabalho foi publicado utilizando o **Galoá** proceedings

das variáveis que não são relevantes. Essa informação pode nutrir análises para otimizar as aferições de qualidade, aumentando a eficácia do processo e reduzindo custos de produção. Assim, mesmo na porção 60-40%, que fornece a menor fração de dados para treino, no banco de dados OXY 51,58% das variáveis não foram utilizadas para compor subconjuntos de preditores. Quando foi utilizada a porção que oferece mais dados ao treino, 90-10%, a identificação das variáveis mais importantes torna-se mais precisa, o que levou ao apontamento de mais variáveis a serem descartadas como preditores. No resultado menos expressivo, 2,08% das variáveis não são relevantes, um resultado que, apesar de ser mais modesto, caracteriza uma informação importante para otimização do processo de qualidade. Como os preditores são relacionados a uma classe que indica a conformidade de qualidade, é possível utilizar essa informação para investigar se as variáveis tidas como não importantes estão sendo utilizadas em outras etapas do processo produtivo, e, se não estiverem sendo utilizadas, podem ser descartadas do processo de aferição da qualidade, tornando a atividade mais eficaz e reduzindo custos.

4 Conclusões

O processo de seleção de variáveis tem papel fundamental para criação de modelos preditivos mais acurados e parcimoniosos, como relatam Guyon e Elisseeff (2003). O modelo proposto, SVD-GMDH, apresenta uma inovação ao implementar um novo ranking de importância das variáveis, baseado em SVD, que elenca as variáveis com base no poder de discriminação das observações. O ranking orienta, de forma iterativa, a inclusão das variáveis em subconjuntos de preditores utilizados no método de classificação Ensamble Logistic GMDH. Esse método utiliza a estratégia *stacking ensemble* para combinar múltiplos algoritmos de aprendizagem de máquina. O SVD-GMDH mostrou-se eficiente para definir o subconjunto ótimo ou subótimo de predição quando comparado a outros modelos para seleção de variáveis, como Anzanello *et al*. (2009), Regressão Logística *Stepwise* e Logistic GMDH-NN.

Um dos primeiros benefícios que pode ser obtido com a aplicação do SVD-GMDH é a melhora da acurácia na classificação de qualidade das amostras entre conforme e não conforme. No conjunto de dados relacionados ao Biodiesel e Diesel brasileiros, classificações incorretas podem levar ao mercado um produto prejudicial ao consumidor. Como consequência, um produto não conforme classificado de forma errada como conforme pode gerar desde um maior consumo de combustível até depreciação acelerada ou falha no motor. Para compreender o impacto de um produto não conforme no mercado, deve-se considerar que, somente no Brasil, há 2 milhões de caminhões ativos, conforme dados da Revista Exame (2019). Ademais, outro fato importante, mas visto de forma indireta, está na imagem da marca do Biodiesel e Diesel brasileiro, que pode, inclusive, ecoar negativamente na esfera internacional.

Com relação aos processos industriais, além da melhora na precisão da classificação, o SVD-GMDH, ao reduzir os subconjuntos de preditores importantes, também aponta um conjunto maior de variáveis não importantes. Nesse sentido, o método coopera para gerar informações que podem ser utilizadas na análise do processo de aferição da qualidade, tendo como consequências positivas a maior eficácia e redução de custos.

Sobre a estrutura do método, um aspecto de destaque é a utilização do algoritmo indutivo GMDH. A heurística GMDH processa automaticamente inter-relações nos dados, selecionar uma estrutura ideal de modelo ou rede e aumentar a precisão dos algoritmos existentes. O GMDH aplicado em redes neurais consegue definir automaticamente a quantidade de camadas, neurônios e quais são as variáveis de entrada necessárias (excluindo as desnecessárias), o que elimina a necessidade de definir os parâmetros do modelo arbitrariamente.

Cabe ressaltar que, mesmo apresentando os melhores resultados, o método torna-se pouco parcimonioso ao envolver diversos algoritmos. Nesse contexto, caberiam estudos futuros sobre a integração do SVD com o GMDH no sentido de melhorar a heurística do GMDH, reduzindo a complexidade do método.

Galoá { Este trabalho foi publicado utilizando o Galoá proceedings

5 Referências

AHMDI, M.H.; Ramezanizadeh, M.; Alhuyi Nazari, M.; Kheradmand, S.; Shamshirband, S. Carbon Dioxide Emission Prediction of Four CIS Countries by Applying a Correlation and GMDH Artificial Neural Network. *Preprints 2019*, 2019. ISSN 2019060227. doi 10.20944/preprints201906.0227.v1. <https://www.preprints.org/manuscript/201906.0227/v1>.

ANZANELLO, M. J.; ALBIN, S. L.; CHAOVALITWONGSE, W. A. Selecting the best variables for classifying production batches into two quality levels. Chemometrics and Intelligent Laboratory Systems, v. 97, n. 2, p. 111–117, 2009. ISSN 0169-7439.

ANZANELLO, M. J. *et al.* PLS-DA wavenumber selection for the categorization of medicine samples based on multiple criteria. Forensic Science International, v. 242, n. Supplement C, p. 111–116, 2014. ISSN 0379-0738. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0379073814002709>.

ANZANELLO, M. J. *et al.* A multivariate-based wavenumber selection method for classifying medicines into authentic or counterfeit classes. Journal of Pharmaceutical and Biomedical Analysis, v. 83, n. Supplement C, p. 209–214, 2013. ISSN 0731-7085. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0731708513001970>.

BRASIL. Lei nº 8.078, de 11 de setembro de 1990. Institui o Código Civil. Presidência da República Casa Civil: Subchefia para Assuntos Jurídicos, Brasília, DF, ano 139, n. 8, p. 1-74, 1990. Disponível em: <http://www.planalto.gov.br/ccivil_03/leis/l8078.html>. Acesso em: 14 jan. 2020.

DAG, O.; KARABULUT, E.; ALPAR, R. GMDH2: Binary Classification via GMDH-Type Neural Network Algorithms—R Package and Web-Based Tool. [S.l.], 2019. v. 12, n. 2, 649 p. Disponível em: <https://cran.r-project.org/package=GMDH2>.

EXAME, O Brasil tem caminhões em excesso – e terá ainda mais. *Revista Exame*. 2019. Disponível em <https://exame.abril.com.br/economia/o-brasil-tem-caminhoes-em-excesso-e-tera-ainda-mais/>. Acessado em 25/06/2020.

FERNÁNDEZ, F. H.; LOZANO, F. H. Gmdh algorithm implemented in the intelligent identification of a bioprocess. In: ABCM Symposium Series in Mechatronics.–2010.–4.–P. [S.l.: s.n.], 2010. p. 278–287.

FERRÃO, M. F. *et al.* Simultaneous determination of quality parameters of biodiesel/diesel blends using HATR-FTIR spectra and PLS, iPLS or siPLS regressions. Fuel, v. 90, n. 2, p. 701–706, feb 2011. ISSN 00162361.

FUKUNAGA, K. Introduction to statistical pattern recognition. [S.l.]: Academic Press, 1990. 591 p. ISBN 9780080478654.

GAUCHI, J.-P.; CHAGNON, P. Comparison of selection methods of explanatory variables in PLS regression with application to manufacturing process data. Chemometrics and Intelligent Laboratory Systems, v. 58, n. 2, p. 171–193, oct 2001. ISSN 01697439. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0169743901001587>.

GOLUB, G.; KAHAN, W. Calculating the Singular Values and Pseudo-Inverse of a Matrix. Journal of the Society for Industrial and Applied Mathematics: Series B, Numerical Analysis, Society for Industrial and Applied Mathematics, v. 2, n. 2, p. 205–224, 1965. ISSN 0887459X. Disponível em: <http://www.jstor.org/stable/2949777>.

Galoá { Este trabalho foi publicado utilizando o **Galoá** proceedings

HUANG, Q. *et al*. Exploiting local coherent patterns for unsupervised feature ranking. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), v. 41, n. 6, p. 1471–1482, 2011. ISSN 1083-4419.

HUNTER, A. Feature Selection Using Probabilistic Neural Networks. Neural Computing & Applications, Springer-Verlag London Limited, v. 9, n. 2, p. 124–132, jul 2000. ISSN 0941-0643. Disponível em: <http://link.springer.com/10.1007/s005210070023>.

IVAKHNENKO, A. G. Heuristic self-organization in problems of engineering cybernetics. Automatica, Elsevier, v. 6, n. 2, p. 207–219, 1970.

JENSEN, R.; SHEN, Q. Computational intelligence and feature selection : rough and fuzzy approaches. [S.l.]: IEEE Press, 2008. 339 p. ISBN 9780470377888.

KHAIRE, U. M.; DHANALAKSHMI, R. Stability of feature selection algorithm: A review. [S.l.]: King Saud bin Abdulaziz University, 2019.

KIRA, K.; RENDELL, L. A. The feature selection problem: Traditional methods and a new algorithm. In: Proceedings of the Tenth National Conference on Artificial Intelligence. AAAI Press, 1992. (AAAI'92), p. 129–134. ISBN 0-262-51063-4. Disponível em: <http://dl.acm.org/citation.cfm?id=1867135.1867155>.

KONDO, T. GMDH neural network algorithm using the heuristic self-organization method and its application to the pattern identification problem. In: Proceedings of the 37th SICE Annual Conference. International Session Papers. Soc. Instrum. Control Eng, 1998. p. 1143–1148. Disponível em: <http://ieeexplore.ieee.org/document/742993/>.

KONDO, T.; PANDYA, A.; ZURADA, J. Logistic GMDH-type neural networks and their application to the identification of the X-ray film characteristic curve. In: IEEE SMC'99 Conference Proceedings. 1999 IEEE International Conference on Systems, Man, and Cybernetics (Cat. No.99CH37028). IEEE, 1999. v. 1, p. 437–442. ISBN 0-7803-5731-0. Disponível em: <http://ieeexplore.ieee.org/document/814131/>.

MASOUMNEZHAD, M.; JAMALI, A.; NARIMAN-ZADEH, N. Robust GMDH-type neural network with unscented Kalman filter for non-linear systems. Transactions of the Institute of Measurement and Control, SAGE Publications Sage UK: London, England, v. 38, n. 8, p. 992–1003, 2016.

MOOSAVIAN, A. *et al*. Comparison of Two Classifiers; K-Nearest Neighbor and Artificial Neural Network, for Fault Diagnosis on a Main Engine Journal-Bearing. Shock and Vibration, v. 20, n. 2, p. 263–272, 2013. ISSN 1070-9622. Disponível em: <http://www.hindawi.com/journals/sv/2013/360236/>.

PASQUALINI, M. *Experience Economy, Brands and Leadership: Reframe Brands, Marketing and Business*. [S.l.]: Amazon Digital Services LLC - Kdp Print Us, 2019. Shared Perspective Series. ISBN 9781094698755.

R CORE TEAM. R: A Language and Environment for Statistical Computing. Vienna, Austria, 2019. Disponível em: <https://www.r-project.org/>.

SKILLICORN, D. B. Understanding complex datasets : data mining with matrix decompositions. [S.l.]: Chapman & Hall/CRC Press, 2007. 236 p. ISBN 1584888326.

VOLTERRA, V. Theory of Functionals and of Integral and Integro-differential Equations. [S.l.: s.n.], 1930. ISBN 0486442845.

XIAO, J.; HE, C.; WANG, S. A classifier ensemble model based on gmdh-type neural network for customer targeting. In: . [s.n.], 2014. p. 259–269. Disponível em: <http://link.springer.com/10.1007/978-3-642-40078-0_22>.

Galoá { Este trabalho foi publicado utilizando o **Galoá** proceedings