CONGRESSO BRASILEIRO
DE ENGENHARIA QUÍMICA EM
INICIAÇÃO CIENTÍFICA

21-24 Julho de 2019
Uberlândia/MG

PREVISÃO DA DEMANDA BIOQUÍMICA DE OXIGÊNIO POR MEIO DE VARIÁVEIS LIMNOLÓGICAS UTILIZANDO MODELOS DE APRENDIZADO DE MÁQUINA EM PYTHON

L. O. M. DA SILVA[1], D. L. C. DA SILVA[1], L. A. VANDERLEI[1], M. C. O. L. FILHO[1] e F. O. CARVALHO[1]

[1] Universidade Federal de Alagoas, Centro de Tecnologia
E-mail para contato: lucas.mendes@ctec.ufal.br

RESUMO – A Demanda Bioquímica de Oxigênio (DBO) é a quantidade de oxigênio indispensável para oxidação de matéria orgânica presente na água e, quando em excesso, pode causar problemas ambientais graves. Dessa forma, o monitoramento de parâmetros limnológicos como a DBO se torna necessário para manutenção da qualidade de águas superficiais. O entendimento da dinâmica de ambientes aquáticos pode ser uma tarefa difícil devido à alta complexidade envolvida nestes fenômenos, a sazonalidade e efeitos espaciais. Assim, o uso de algoritmos de aprendizado de máquina para o estudo da dinâmica em ambientes aquáticos vem sendo cada vez mais recorrente. Reconhecendo a demanda surgida, este trabalho buscou aplicar *Support Vector Machine* (SVM) e Redes Neurais Artificiais (RNA), técnicas de aprendizado de máquina, para quantificar a DBO no Rio Lam Tsuen, situado na China, a partir de variáveis limnológicas.

1. INTRODUÇÃO

Devido à água ser tratada com negligência, este recurso que costumava ser apontado como abundante na natureza começa a sofrer com o uso e ocupação do solo, resultando em crises hídricas cada vez mais marcantes. Na busca de uma melhor gestão hídrica, acompanhar as variáveis limnológicas, parâmetros físico-químicos e biológicos do meio, como a Demanda Bioquímica de Oxigênio, é importante para o monitoramento da qualidade dos corpos hídricos e para o entendimento da dinâmica destes ambientes aquáticos (Júnior *et al*, 2019).

No entanto, tal entendimento pode ser uma tarefa difícil devido à alta complexidade envolvida nestes fenômenos. Deste modo, modelos matemáticos de fundamentação teórica tendem a ser mais complexos nestes problemas e muitas vezes podem requerer considerações teóricas para resolução que podem não condizer com a realidade. Assim, os modelos empíricos apresentam alto potencial de aplicação no estudo dinâmico de corpos hídricos.

O uso de algoritmos de aprendizado de máquina para o estudo da dinâmica em ambientes aquáticos é bastante recorrente. Sarkar e Pandey (2015) fizeram uso de RNA, uma técnica de aprendizado de máquina baseada na forma como o cérebro humano aprende, para estimar a concentração de Oxigênio Dissolvido no Rio Yamuna na Índia. Já Luiz *et al* (2018) aplicaram

a SVM, uma técnica de aprendizado estatístico, para previsão de níveis diários em fontes de água subterrânea no aquífero livre, situado no oeste do Rio Grande do Sul.

Conforme a PYPL, plataforma de controle de versão de códigos fonte do GITHUB, a linguagem Python foi a segunda mais popular no ano de 2018. Isso se deve ao fato de essa linguagem ser gratuita, de fácil aprendizado e possuir bibliotecas que contém diversas técnicas para a elaboração de sensores virtuais baseados em inteligência artificial e estatística.

Sendo assim, reconhecendo o grande potencial de aplicação de algoritmos de aprendizado de máquina e com o intuito de contribuir para o desenvolvimento de técnicas de monitoramento da qualidade de água a partir de modelos empíricos, este trabalho buscou avaliar a aplicabilidade das máquinas de aprendizados (RNA e SVM) em Python, utilizando a biblioteca *scikit-learn*, para a previsão da DBO no Rio Lam Tsuen, na China, com uso de variáveis limnológicas como variáveis de entrada. Para a avaliação dos modelos, utilizou-se os parâmetros estatísticos *root-mean-square error* (RMSE) e coeficiente de determinação (R^2), medidas frequentemente usadas para avaliar os valores preditos em relação aos reais.

2. MATERIAIS E MÉTODOS

2.1. Dados Utilizados

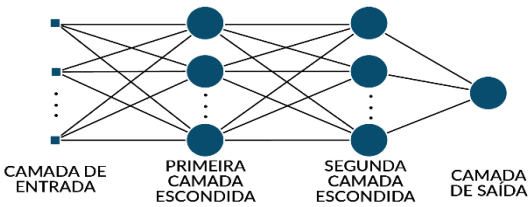
A partir de pesquisas realizadas nos bancos de dados nacionais e estrangeiros para obter um conjunto de dados suficientemente grande de qualidade de águas superficiais, foram escolhidos os dados do Rio Lam Tsuen, rio chinês que atravessa o distrito de Tai Po, uma cidade do sul de Hong Kong na China, visto que o crescimento industrial e populacional acelerado, somado aos períodos de baixos índices pluviométricos agrava a situação dos corpos hídricos. Foram disponibilizadas 288 amostras desse rio, entre 2009 e 2011, que foram coletadas no porto e no canal de Tolo no qual os parâmetros analisados foram: Temperatura, Sólidos suspensos, Sólidos totais, Fósforo, Nitrito, Nitrato, OD ($mg.L^{-1}$ e porcentagem de saturação), Condutividade, Amônia, DQO e DBO. Para a validação dos modelos, a base de dados foi dividida de forma randômica entre treinamento e teste, em que 75% dos valores foram utilizados para o treinamento, e os outros 25%, para o teste.

2.2. Redes Neurais Artificiais (RNA)

Com a capacidade de "aprender" o comportamento de um sistema real complexo de modo similar ao realizado pelo cérebro humano, as RNAs podem ser definidas como algoritmos capazes de processar informações e fazer inferências a partir delas. Conforme Haykin (2008), a Figura 1 mostra o diagrama de uma *Multilayer Perceptron* (MLP) com duas camadas ocultas e uma camada de saída. Com uma rede totalmente conectada, esse tipo de arquitetura da rede faz com que o sinal flua em direção para frente, da esquerda para a direita por meio dos neurônios artificiais, os quais se conectam via interligações denominadas sinapse. Cada neurônio possui uma função de ativação, a qual limita a amplitude do sinal recebido pela sinapses, que pondera dos dados de entrada por meio de pesos sinápticos.

Dentre as várias formas de aprendizagem de uma rede neural, destaca-se o aprendizado supervisionado, cuja principal característica é a disponibilidade de dados do sistema para formar padrões de entrada-saída. Com isso, no conjunto de treinamento, a saída prevista pela rede é comparada com a saída esperada e, a partir do sinal de erro gerado, os pesos sinápticos são ajustados, possibilitando a rede inferir valores não utilizados na etapa de treino. O algoritmo de treinamento utilizado para o ajuste do pesos foi o LBFGS (do inglês, *Limited-memory Broyden–Fletcher–Goldfarb–Shanno*), um método de otimização pertencente à família dos métodos Quasi-Newton.

Figura 1 – Diagrama de uma MLP com duas camadas ocultas e uma camada de saída.



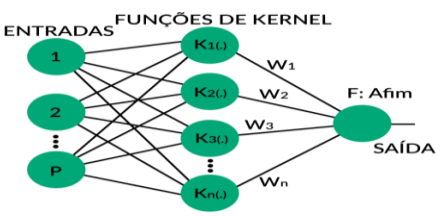
Fonte: Autores (2019).

2.3. *Support Vector Machine* (SVM)

Desenvolvida por Vapnik, a SVM é uma técnica de aprendizado de máquina (AM) supervisionado baseada na Teoria de Aprendizagem Estatística. Neste tipo de AM um conjunto de dados de treinamento contendo dados de entrada com suas respectivas saídas desejadas é apresentado à máquina. O objetivo é que, a partir desses dados, a máquina produza um classificador capaz de indicar saídas corretas quando novas entradas forem apresentadas (Lorena, Carvalho, 2003).

O método é utilizado na classificação e regressão de dados para conjuntos linearmente ou não-linearmente separáveis. Para dados linearmente separáveis, o processo consiste na escolha de um hiperplano que deve ser ajustado por meio dos dados de treinamento. Nos casos onde os dados não são linearmente separáveis, transformações não-lineares são realizadas utilizando as funções de Kernel, as quais permitem a migração dos dados para uma dimensão maior, onde será mais provável que sejam linearmente separáveis (Gonçalves, 2016). Segundo Lorena e Carvalho (2003), após realizadas essas transformações é possível aplicar a SVM linear e obter um hiperplano que garanta uma boa generalização para o classificador. A Figura 2 ilustra um fluxograma representativo com a topologia geral de uma SVM.

Figura 2: Arquitetura da SVM



Fonte: Autores (2019).

3. RESULTADOS

3.1. Modelo RNA

A Tabela 1 mostra os resultados observados para os modelos desenvolvidos através da utilização da RNA para inferência do parâmetro físico-químico DBO. O melhor resultado observado está destacado em negrito.

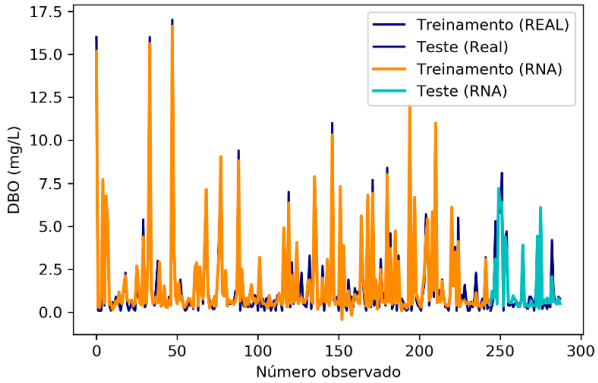
Tabela 1- Índices estatísticos RMSE e R^2 para as arquiteturas de RNA testadas.

Algoritmo de Treinamento	**Camada Escondida**		**Treinamento**		**Teste**	
	Função de ativação	**Nº de neurônios**	R2	RMSE	R2	RMSE
LBFGS	**tanh** $f(x) = \frac{2}{(1+e^{-2x})-1}$	7	0,9374	0,7029	0,7928	0,8955
		8	**0,9602**	**0,5351**	**0,8019**	**0,8756**
		9	0,9436	0,5384	0,6441	1,3265
	Sigmoidal $f(x) = \frac{1}{(1+e^{-x})}$	7	0,7916	1,2227	0,7894	0,9748
		8	0,8570	1,0147	0,7129	1,0541
		9	0,9095	0,8070	0,6361	1,1869
	relu $f(x) = \max(0, x)$	7	0,8891	1,1380	0,3910	1,5353
		8	0,8716	1,5290	0,5967	1,0049
		9	0,9575	0,5532	0,2356	1,7201

Fonte: Autores (2019).

Foi observado que os melhores resultados estavam para arquiteturas com de 7 a 9 neurônios na camada escondida. Abaixo disto, dentre as funções de ativação utilizadas, foi visto que a tangente hiperbólica foi a que mais se adequou ao problema. A regressão do melhor modelo obtido pode ser vista na figura 3.

Figura 3 - Regressão para o melhor modelo entre os testados.



Fonte: Autores (2019).

3.2. Modelo SVM

Para o modelo da SVM, a tabela abaixo apresenta os resultados estatísticos obtidos, bem como os parâmetros, escolhidos de forma heurística, utilizados para cada função Kernel.

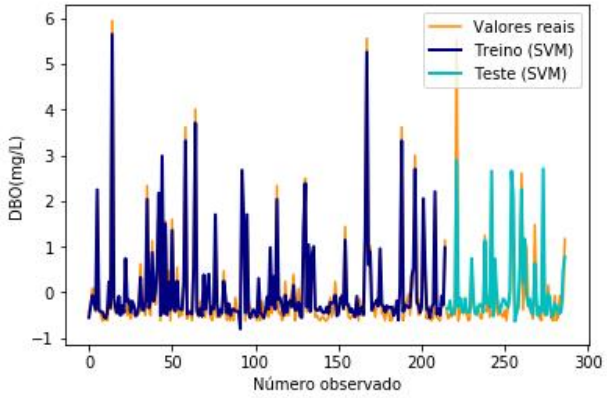
Tabela 2 – Índices estatísticos para o modelo SVM para cada função Kernel testada.

Função de Kernel	Parâmetros	Treino	Teste
Linear $\varphi(x) = x_i^T x_j$	$C = 10$, $\varepsilon = 0.01$	$R^2 = 0.6505$ RMSE = 0.6002	$R^2 = 0.5134$ RMSE = 0.6638
Sigmoide $\varphi(x) = tanh(\gamma \cdot x_i^T x_j + k)$	$C = 10^5, \varepsilon = 10^{-6}$ $\gamma = 10^{-5}, k = 0.054$	$R^2 = 0.6520$ RMSE = 0.5989	$R^2 = 0.5149$ RMSE = 0.6627
RBF $\varphi(x) = exp\left(-\frac{\left\|\|x_i - x_j\right\|^2}{2\sigma^2}\right)$	$\boldsymbol{C = 10^5, \sigma = 0.07,}$ $\boldsymbol{\varepsilon = 0.3}$	**$R^2 = 0.9613$ RMSE = 0.1997**	**$R^2 = 0.7123$ RMSE = 0.5104**

Fonte: Autores (2019)

Os modelos utilizando as funções Kernel linear e sigmoide apresentaram um pior desempenho, enquanto que a RBF mostrou o desempenho mais satisfatório. Além disso, vale salientar que os parâmetros de cada função foram escolhidos de forma heurística, levando em consideração que valores muito alto do fator de penalidade (C) ou muito baixos da margem de tolerância (ε) podem causar problemas de super-ajuste, situação na qual o modelo se ajusta muito bem aos dados de treinamento, mas não generaliza bem no teste. A figura 4 mostra o melhor modelo obtido pelo modelo SVM. Maiores detalhes sobre as funções Kernel utilizadas podem ser encontrados em Gonçalves (2016).

Figura 4– Regressão utilizando a função Kernel RBF



Fonte: Autores (2019).

4. CONCLUSÃO

A RNA apresentou um desempenho ligeiramente superior em relação à SVM para previsão da DBO no Rio Lam Tsuen. Vale destaca que o mesmo conjunto de dados foi utilizado em outros softwares comerciais e foram obtidos melhores resultados, indicando que a implementação da SVM e da RNA em Python precisa ser melhor avaliada, o que pode ser feito explorando outras bibliotecas de aprendizado de máquina além da *scikit-learn* – biblioteca utilizada neste trabalho. Sendo assim, dada a natureza empírica dos modelos, a abordagem utilizada neste trabalho para inferir a DBO a partir de variáveis limnológicas pode ser considerada adequada para solução de problemas de regressão.

5. REFERÊNCIAS

GONÇALVES, A.R. **Máquina de Vetores Suporte**. 2016. Disponível em: <https://www-users.cs.umn.edu/~agoncalv/arquivos/pdfs/svm.pdf>. Acesso em: 27 de Fevereiro de 2019.

HAYKIN, S. Multilayer Perceptrons. In: **Neural Networks and Learning Machines**. 3rd ed. New York: Prentice Hall, 2008. cap. 4, p. 123-221.

JÚNIOR, A.P.; CONCEIÇÃO, C.S.; LOBO, R.R.; SANTOS, C.O.R. SARDINHA, A.S. Association between ephemeroter, plecoptera and trichoptera and the limnimetric parameters of the water quality index. *Revista Brazilian Applied Science Review*. v. 3, n. 2, p. 839-863. Curitiba,. 2019.

LORENA, Ana C.; CARVALHO, André C. P. L. F. **Introdução às Máquinas de Vetores Suporte**. Instituto de Ciências Matemáticas e de Computação - USP. São Carlos, 2003

LUIZ, T.B.P; GAIARDO,G.F.;SILVA,J.L.S. Utilização de máquina de vetor de suporte para previsão de níveis de água subterrânea. Revista Águas Subterrâneas. V. 32, n. 1, 2018.

SARKAR, A; PANDEY, P. River Water Quality Modelling Using Artificial Neural Network Technique. *International conference on water resources, coastal and ocean engineering (icwrcoe 2015).* V.4, p.1070-1077, 2015.