AVALIAÇÃO DE SISTEMAS INTELIGENTES NO DESENVOLVIMENTO DE SENSORES VIRTUAIS PARA DETERMINAÇÃO DA QUALIDADE DE ÁGUA

M.C.de O. LIMA FILHO[1], L. C. BARRETO[2], B. C. V. dos SANTOS[3]; W. B. A. de M. SILVA[4] e F. O. CARVALHO[5]

[1] Universidade Federal de Alagoas, Centro de Tecnologia, Departamento de Engenharia Química

[2] Universidade Federal de Alagoas, Centro de Tecnologia, Departamento de Engenharia Química

[3] Universidade Federal de Alagoas, Centro de Tecnologia, Departamento de Engenharia Civil

[4] Universidade Federal de Alagoas, Centro de Tecnologia, Departamento de Engenharia Química

5 Universidade Federal de Alagoas, Centro de Tecnologia, Departamento de Engenharia Química

E-mail para contato: marcone_oliv@hotmail.com

RESUMO – A água é um recurso natural abundante, entretanto a má utilização e gestão deste recurso faz com que a quantidade e qualidade seja diminuída constantemente. Assim, é importante cautela para conservar sua qualidade. Quando utilizada de forma imprudente, há alteração da qualidade química de reservas. Para garantir a integridade dos corpos hídricos surgem estratégias de monitoramento dos parâmetros físico-químicos da água. Nessa perspectiva, entender a dinâmica dos ambientes aquáticos por meio da avaliação do comportamento das variáveis limnológicas é importante para fornecer informações sobre o comprometimento da qualidade da água frente aos impactos sofridos. Contudo, o acompanhamento pode se tornar uma atividade laboriosa quando se usam modelos fenomenológicos. Assim, este trabalho buscou avaliar o desempenho de duas técnicas empíricas de regressão, a Rede Neural Artificial (RNA) e a Support Vector Machine (SVM), aplicando-as no desenvolvimento de um sensor virtual para inferência da concentração de clorofila, variável de difícil medição imediata, na água da represa de Guarapiranga, São Paulo, Brasil, através de parâmetros físico-químicos, como pH, temperatura, condutividade, oxigênio dissolvido (OD), potencial de oxirredução (ORP) e profundidade. O modelo da RNA com algoritmo de treinamento de Regularização Baysiana (RB) apresentou melhor desempenho, com R^2 iguais a 0,94 e 0,85 e MSE iguais a 23,67 e 53,81 nas etapas de treino e teste respectivamente.

PALAVRAS-CHAVE: Qualidade de água, sensor virtual, rede neural artificial, *support vector machine.*

1. INTRODUÇÃO

Embora a água seja um recurso natural abundante, as grandes demandas acompanhadas pela crescente deterioração dos recursos hídricos pelas múltiplas atividades humanas vêm alterando a qualidade da água das reservas superficiais e subterrâneas, e isso faz com que seja necessária a elaboração de estratégias para obtenção de modelos para a qualidade da água. Contudo, uma vez que as fontes superficiais de água estão mais acessíveis que os aquíferos subterrâneos, estas se mostram mais suscetíveis à contaminação e por isso será o foco abordado no presente trabalho. Moreira *et al* (2012) avaliou águas superficiais em Minas Gerais e constatou uma alta contaminação por ação de agrotóxicos, destacando a importância de monitoramento destas regiões de maior impacto. Outros fatores contaminantes que atingem os sistemas hídricos podem ser citados: os descartes domésticos, industriais e as mudanças químicas causadas por substâncias tóxicas. O entendimento da dinâmica dos ambientes aquáticos por meio da avaliação do comportamento das variáveis limnológicas (físicos, químicos e biológicos da água) é um fator importante para fornecer informações relativas ao grau de comprometimento da qualidade da água frente aos impactos sofridos (Fernandes e Gomes, 2016).

Nesse processo de monitoramento da qualidade da água, os índices utilizados que tem como objetivo retratar a qualidade da água nos pontos de interesse resultam em dados multidimensionais não lineares, demandando muito tempo para a análise dos resultados, além de um alto custo computacional para a sua obtenção. A fim de sanar tais dificuldades é recorrente o uso de métodos que forneçam uma modelagem e avaliação dos dados a custos mais baixos relacionados ao tempo e processamento, sendo os mais usuais os modelos empíricos e fenomenológicos, contudo, devido às limitações dos modelos fenomenológicos e dada a complexidade dos dados obtidos é recorrente o uso de modelos empíricos, que além de se adequar bem aos tipos de dados obtidos, apresentam uma crescente disponibilidade dos métodos estatísticos e de sistemas inteligentes para modelagem empírica de dados, tornando possível a determinação de padrões e relações entre as variáveis (Parracho, 2010).

Para o desenvolvimento dos modelos empíricos para problemas de previsão, diversos autores como Khosravi *et al* (2018) e Lafetá et al (2018) reportam o uso de algoritmos de aprendizado de máquina, dentre os quais pode-se destacar as Redes Neurais Artificiais (RNA) e os *Support Vector Machine (SVM)*, os quais são capazes de reconhecer padrões, interpretar, descrever processos através da formulação de modelos empíricos, que se adequam bem a fenômenos naturais de alta complexidade. Desta forma, neste trabalho buscou-se avaliação da utilização dessas duas técnicas na modelagem da qualidade de agua através do desenvolvimento de sensores virtuais para inferenciar a concentração de clorofila em

amostras de água da represa de Guarapiranga - SP, Brasil, efetuando uma comparação do desempenho das duas técnicas.

2. MATERIAIS E MÉTODOS

2.1. Redes Neurais Artificiais

As Redes Neurais Artificiais (RNA) são conjuntos de unidades de processamento baseadas no funcionamento do cérebro humano, reproduzindo o seu sistema organizacional de informação (Haykin, 2001). Cada unidade do conjunto é um neurônio artificial que por sua vez imita o processamento realizado pelo neurônio biológico. Assim como o cérebro necessita de estímulos para adquirir conhecimento, a RNA aprende a partir de experiências anteriores e quanto maior o número de exemplos (dados), maior a capacidade de generalizar a informação. O processo de aprender a partir de situações em que se conhece um conjunto de dados composto por entradas e suas respectivas saídas é denominado aprendizado supervisionado (Camelo, 2017)

As RNA's são modelos empíricos amplamente utilizados para ajuste de dados devido a sua alta capacidade de processamento paralelo, habilidade de generalização e previsão de padrões. Dentre os inúmeros tipos de RNA que podem ser utilizadas para regressão e ajuste de dados, a mais utilizada é a rede Perceptron de Multi Camadas (*Multi Layer Perceptron)*, apresentada na Figura 1. Esta possui uma arquitetura com três camadas: camada de entrada, camada intermediária e a camada de saída.

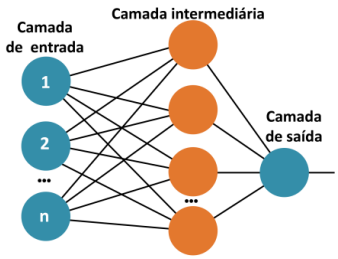

Figura 1 – Rede MLP.

Os dados de entrada em cada neurônio sofrem uma ponderação através dos respectivos pesos sinápticos. Além dos pesos, também há introdução dos bias, parâmetros que também são ajustados no processo de aprendizado. Sendo assim, a etapa de aprendizado ou treinamento, trata-se de um problema de otimização matemática no qual os pesos sinápticos e bias são ajustados com objetivo de formular um modelo empírico robusto capaz de generalizar a informação obtida para produzir saídas que apresentem valores o mais próximo

possível do valor real. O métodos de Levenberg-Marquardt e Regularizaçao Baysiana costumam ser bastante utilizados para efetuar essa otimização (Pinheiro *et al*, 2017).

2.2. Support Vector Machine

As Máquinas de Vetores Suporte (*Support Vector Mchine) (SVM)* são algoritmos de aprendizado de máquina desenvolvidos por Vapnik *et al* (1999) e tem embasamento teórico na Teoria de Aprendizado Estatístico e Otimização Matemática. Devido a sua teoria bem definida, além de fatores como ótima capacidade de generalização, robustez para problemas de grandes dimensões e também garantia da convergência, as SVM's têm sido recorrentes na solução de problemas como previsão de uma variável resposta (Lorena e Carvalho, 2007). A SVM foi inicialmente desenvolvida para solucionar problemas de classificação em um conjunto de dados linearmente separáveis. O algoritmo tem então como princípio de funcionamento definir um hiperplano no espaço que consiga separar os dados em duas regiões distintas, ou seja, duas classes diferentes. Inicialmente, o hiperplano é definido a partir da Equação 1.

$$f(x) = w.x + b \tag{1}$$

Onde w é o vetor peso, b é o vetor denominado de *bia.* Inúmeros hiperplanos podem ser definidos porém, o objetivo é obter aquele que melhor separe os dados, um hiperplano ótimo. Deste modo, por meio de um aprendizado supervisionado, ou seja, a partir do conjunto de entradas e saídas, os parâmetros são ajustados para obter o hiperplano ótimo. Para determinar o hiperplano, a SVM faz uso dos pontos mais próximos do hiperplano e a partir da distância entre o hiperplano e os primeiros pontos de cada classe define-se uma margem de separação. A Figura 2 ilustra o princípio do funcionamento de uma SVM (Whan e Wu, 2015) .

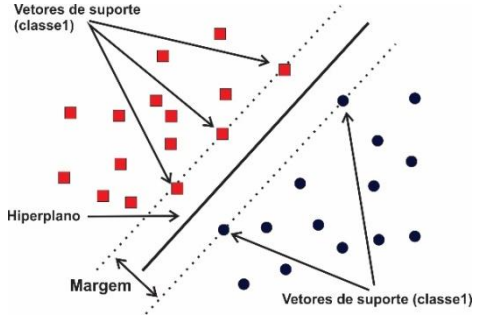

Figura 2 - Hiperplano para uma SVM.

A partir daí, o problema se resume a definir um hiperplano através da maximização ou minimização da margem. Quando o objetivo é realizar separação dos dados em classes

distintas, a busca é pela maximização da margem, já quando o objetivo é realizar uma inferência (regressão), a busca se dá pela minimização ajustando então o hiperplano ótimo ao conjunto de dados. Quando o conjunto de dados não é linearmente separável, a SVM utiliza então uma função de Kernel, a qual é capaz de transportar os dados para um espaço característico de maior dimensão, conforme pode ser observado na Figura 3. Desta forma, os dados que antes não podiam ser separados por um plano passam a ser separáveis. Segundo Lorena e Carvalho (2007), as funções Kernel mais utilizadas são a radial basis (RBF),gaussiana, polynomial e função sigmoid.

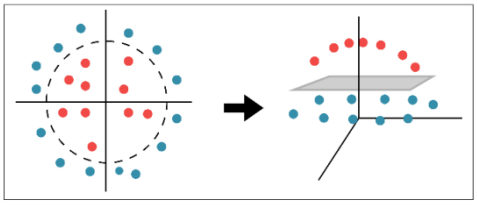

Figura 3 - Função de Kernel mapeando um conjunto não-linearmente separável.

2.3. Aquisição de Dados e Desenvolvimento dos Sensores Virtuais

Realizaram-se pesquisas em diversos bancos de dados do Brasil, desta forma, foi possível reunir dados referentes à represa de Guarapiranga, a qual compõe o segundo maior sistema de produção de água potável do estado de São Paulo. Os principais parâmetros analisados nesse conjunto de dados foram a condutividade, oxigênio dissolvido (OD), pH, potencial de oxirredução (ORP), temperatura, turbidez e clorofila (mg/L). Através da regressão e análise dos dados, buscou-se o desenvolvimento de sensores virtuais, utilizando as variáveis de entrada (condutividade, oxigênio dissolvido (OD), pH, potencial de oxi-redução (ORP), temperatura, turbidez e profundidade) para inferênciar a concentração de clorofila na represa, efetuando a comparação entre o desempenho da RNA e da SVM, utilizou-se a plataforma Matlab ® para o desenvolvimento dos modelos regressores. Dentre os parâmetros ajustáveis da RNA, variou-se o número de neurônios na camada intermediária e o algoritmo de otimização já para a SVM variou-se o tipo de função Kernel, avaliando a melhor topologia das duas técnicas. Os critérios estatísticos utilizados para avaliação dos modelos foi o erro quadrado médio (MSE) e o coeficiente de determinação (R^2).

3. RESULTADOS E DISCUSSÃO

3.1. Redes Neurais Artificiais

A Tabela 1 mostra os resultados obtidos para os modelos desenvolvidos através da utilização da RNA, ressaltando a topologia de rede que apresentou o melhor desempenho no desenvolvimento do sensor virtual para inferência da concentração da clorofila.

Tabela 1 - Índices estatísticos MSE e R² para as topologias de RNA testadas.

Número de Neurônios	Algoritmo de treinamento							
	Levenberg-Marquardt				Regularização Bayesiana			
	R^2		MSE		R^2		MSE	
	Treinamento	teste	treinamento	teste	Treinamento	teste	treinamento	teste
7	0,8665	0,7692	49,20	89,01	0,8745	0,7674	51,03	85,05
8	0,8059	0,8147	60,65	69,97	0,8941	0,7856	43,48	73,25
9	0,8721	0,7916	52,05	77,35	0,8991	0,7732	40,86	92,62
10	0,8703	0,8319	50,99	74,56	**0,9437**	**0,8502**	**23,67**	**53,81**

Ao aumentar o número de neurônios da camada intermediária foi observado uma melhora no desempenho da rede, iniciando com 7 neurônios (quantidade de variáveis de entrada) até 10 neurônios (melhor resultado), já que a partir de 11 neurônios foi observado que o desempenho diminui bastante indicando que poderia está ocorrendo um sobre ajuste. Portanto, a melhor arquitetura observada heuristicamente foi com 10 neurônios na camada intermediária. O algoritmo de treinamento também se mostrou um parâmetro importante a ser avaliado. Regularização Bayesiana demonstrou maior desempenho na predição de clorofila atingindo maiores valores de R^2 e menores valores de MSE tanto na etapa de treino quanto de teste quando comparado ao algoritmo Levenbergue-Marquardt.

3.2. Support Vector Machine

A segunda técnica avaliada foi a *Support Vector Machine*. Através da Tabela 2 é possível visualizar o desempenho dessa técnica na estimativa do parâmetro clorofila através dos valores estáticos (MSE e R^2) apresentados.

Tabela 2 - Índices estatísticos MSE e R² para as topologiasde SVM testadas.

Função de Kernel	R^2	MSE
Linear	0,48	110,12
Quadrática	0,50	106,25
Cúbica	-0,23	261,01
Gaussiana Média	**0,60**	**85,28**

Após testadas quatro funções de Kernel diferentes, é visto que a função gaussiana é a que obteve melhor resultado para predição de clorofila. Lorena e Carvalho (2007) apontaram a gaussiana como uma das funções mais utilizadas nas aplicações de SVM. Entretanto, podemos observar que o desempenho da SVM para o ajuste do conjunto de dados estudados

foi bastante inferior do que o da RNA de modo geral. Para estabelecer uma comparação entre o desempenho das duas técnicas avaliadas frente ao problema estudado, foi elaborada a Tabela 3, a qual compara o desempenho da melhor topologia das duas técnicas.

Tabela 3 – Comparação entre melhores topologias testadas.

Topologia	R	MSE
RNA(10 neurônios e Regularização Bayesiana)	**0,9437**	**23,67**
SVM(Gaussiana Média)	0,60	85,28

Dessa forma, pôde-se verificar que o modelo RNA (10 neurônios e algoritmo de treinamento Regularização Bayesiana) superou a SVM na estimativa da concentração de clorofila na água para o conjunto de dados avaliados. Este resultado pode ser denotado pelo maior valor para R^2, o que indica melhor ajuste dos dados quando se compara o que foi previsto e o valor real da variável para as entradas testadas e um menor MSE, o que indica menor erro acumulado de previsão. A Figura 4 mostra o resultado da regressão para a melhor topologia da RNA.

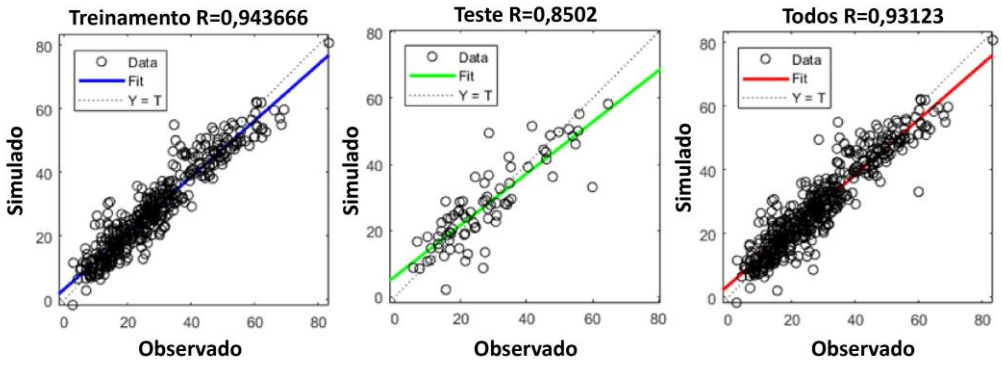

Figura 4 – Regressão para melhor o modelo de RNA.

4. CONCLUSÃO

A partir dos resultados apresentados neste trabalho, ficou evidenciado que a Rede Neural Artificial apresentou um melhor desempenho em relação a *Support Vector Machine* no desenvolvimento de um sensor virtual para o conjunto de dados analisado. Devido aos melhores resultados apresentados nas etapas de treino e teste, podemos ressaltar que há um maior grau de confiabilidade para utilização das RNA's como um sensor virtual para a medição da variável clorofila (analisada neste trabalho) do que com uso de SVM's. Entretanto, no que se refere à utilização da SVM, faz-se necessária um novo teste com a utilização de uma quantidade maior de dados, buscando promover uma maior robustez na etapa de treinamento e posteriormente na etapa de teste.

AGRADECIMENTOS

Os autores agradecem ao Laboratório de Sistemas Inteligentes Aplicados – LABSIA por todo suporte teórico e técnico necessário para a realização deste trabalho.

REFERÊNCIAS

CAMELO, H. N., LUCIO, P.S. , LEAL JÚNIOR, J. V. Modelagem da velocidade do vento usando metodologias ARIMA, HOLT-WINTERS E RNA na previsão de geração eólica no nordeste brasileiro**.** *Rev. Bras. de Clim. V. 21, 2017*

FERNANDES, T. S; GOMES, L. N. L. Avaliação do comportamento de parâmetros limnológicos de qualidade da água na região mais profunda do lago Paranoá/DF. *XIV ENEEAmb & fórum latino americano de engenharia e sustentabilidade*, Brasília, 2016.

HAYKIN, S.S. *Redes Neurais: Princípios e Prática*. 2 ed. Bookman Companhia Ed, 2001.

KHOSRAVI, A; KOURY, R.N.N.; MACHADO, L.; PABON, J.J.G. Prediction of wind speed and wind direction using artificial neural network, support vector regression and adaptive neuro-fuzzy inference system. *Rev. Sust. Ener. Tec. and Assessm.*, v. 25, 2018.

LAFETÁ, B.O.; SANTANA, R.C NOGUEIRA, G.S. NEVES,J.C.L.; PENIDO,T.M. A. Macronutrients use efficiency in eucalypt by non-destructive methods estimated by artificial neural networks. *Ciên. Flor*., v. 28, n. 2, p. 613-623, 2018

LORENA, A. C.; CARVALHO, A. C. de. Uma introdução às support vector machines. *Rer. de Inf. Teó. e Aplic.*, v. 14, n. 2, p. 43–67, 2007.

MOREIRA, J.C.; PERES,F.; SIMÕES,A.C.;PIGNATI,W.A.; DORES,E.C.; VIEIRA,S.N. ; MOTT, T. Groundwater and rainwater contamination by pesticides in an agricultural Region of Mato Grosso State in Central Brazil. Ciên. e Saúde Col., v. 17,p 1557-1568, 2012.

PARRACHO, P. M. V. A. PATTERN. Identificação de padrões em mercados bolsistas. 2010. 99 p. *Dissertação (Mestrado em Engenharia Informática e de Computadores)* - Universidade Técnica de Lisboa, Lisboa, 2010.

PINHEIRO,E.; RUTHER,R.; LOVATO, A. Applicability of levenberg-marquardt algorithm for power generation analysis of the system photovoltaic. Rev. Cient. Elet. De Eng. De Prod. V. 17, n.4., 2017.

WHAN, J.; WU,J. A robust combination approach for short-term wind speed forecasting and analysis e Combination of the ARIMA (Autoregressive Integrated Moving Average), ELM (Extreme Learning Machine), SVM (Support Vector Machine) forecasts using a GPR (Gaussian Process Regression) model. *Energy*. V. 93, p. 41-56, 2015;