Avaliação de Desempenho de Classificadores de Ciclos Hidrológicos em Reservatórios de Água na Região Amazônica

Jean Arouche Freire[1]
Yuri Santos[1]
Jefferson Morais[1]
Terezinha Oliveira[2]

[1]Faculdade de Computação – Universidade Federal do Pará (UFPA)
Rua Augusto Corrêa, nº 1 –Guamá – 66.075-110 – Belém – PA – Brasil

[2]Faculdade de Matemática e Estatística – Universidade Federal do Pará (UFPA)
Rua Augusto Corrêa, nº 1 –Guamá – 66.075-110 – Belém – PA – Brasil

jeanarouche@gmail.com, yuri.nassar@gmail.com, {jmorais,tfo}@ufpa.br

Abstract. *This paper systematically evaluates classifiers in the prediction of hydrological cycles from the change of physico-chemical parameters and metals of water reservoir of the hydroelectric power plant Tucuruí. The methodology initially is to conduct an exploratory analysis of data in order to extract only the most relevant variables of the observed samples. The choice of the parameter values of the classifiers was made with automatic model selection. Results of applying Artificial Neural Networks, K-nearest neighbors, support vector machines and Random Forest techniques are presented. The results indicate that the Random Forest classifier showed the best performance with a percentage rating of 7.8 % of incorrect predictions. These values can be considered significant, since there is a great variability of physico-chemical parameters and metals in the hydrological cycles where are the sampling stations of study area.*

Resumo. *Este artigo tem como objetivo avaliar sistematicamente classificadores na predição dos ciclos hidrológcos a partir da alteração dos parâmetros físico-químicos e metais da água no reservatório da Usina Hidrelétrica de Tucuruí. A metodologia consiste inicialmente em realizar uma análise exploratória dos dados com o objetivo de extrair somente as variáveis mais relevantes das amostras observadas. A escolha dos valores dos parâmetros dos classificadores foram feitas com seleção automática de modelo. Resultados aplicando técnicas de redes neurais artificiais, k-vizinhos mais próximo, máquinas de vetores de suporte e random forest são apresentados. Os resultados obtidos indicam que o classificador random forest foi o que apresentou melhor desempenho com percentual de classificação de 7,8% de predições incorretas. Estes valores podem ser considerados significativos, pois existe uma grande variabilidade dos parâmetros físico-químicos e metais nos ciclos hidrológicos onde se encontram as estações de amostragem da área de estudo.*

1. Introdução

A água é um dos principais recursos naturais do meio ambiente, e em forma potável, é essencial para vida humana, assim como para os ecossistemas aquáticos distintos. Neste

sentido, é primordial que se faça um controle rigoroso da qualidade da água que pode ser afetada por processos naturais, como intemperismo e erosão do solo, assim como pela ação antropogênica [Singh et al. 2009]. Para esse desdobramento, [Gastaldini et al. 2002] chamam atenção para necessidade de se averiguar e diagnosticar elementos que influenciam a qualidade da água em corpos d'água, prevendo impactos futuros oriundos de fenômenos climáticos ou condições específicas.

Para avaliação de possíveis impactos ambientais em corpos d'água, como o reservatório da usina hidrelétrica (UHE) de Tucuruí, provenientes de fenômenos climáticos, em especial ciclos hidrológicos (CHs), é fundamental que se faça um manejo sistemático e que o monitoramento da qualidade da água possa auxiliar na administração dos recursos hídricos conservando os padrões de qualidade que são definidos por legislação [Vrana et al. 2005]. Esses padrões de qualidade podem ser averiguados através da análise dos parâmetros físicos-químicos e metais (PFQM) em corpos de água [Tundisi 1999]. Assim, além da assistência na gestão ambiental em resevatórios, tais expectativas podem também favorecer a preservação e o equilíbrio dos ecossistemas aquáticos.

Dentro desse contexto, [Bittencourt and Amadio 2007] colocam a importância de se classificar CHs, pois através do estudo dessas variações sazonais pode-se fazer estimativas climáticas que contribuam na administração dos recursos hídricos em corpos d'água na região amazônica. Essas perspectivas podem ser potencializadas com o auxílio de recursos computacionais como suporte aos sistemas de gestão ambiental [Bertholdo et al. 2013].

Para [Galvão 1999], apesar da sazonalidade ambiental não ser constante em relação ao tempo, um ponto diferencial para estimativas climáticas utilizando recursos computacionais é o emprego de inteligência computacional (IC). Pois, além de apresentar valores aceitáveis na classificação de fenômenos climáticos, também configura-se como alternativa ou complementos as técnicas consagradas da estatistica, pesquisa operacional e modelagem numérica.

Diante dos pontos levantados seria válido contribuir na avaliação de técnicas de inteligência que melhor se ajustam as estimativas de classificação de CHs, levando em consideração os PFQM da água em reservatórios na região amazônica, em especial o da UHE de Tucuruí.

Existe um acervo de trabalhos científicos relacionados a proposta da pesquisa realizados com o objetivo de identificar e fazer estimativas de variações sazonais climáticas que possam inferir algum impacto em contextos ambientais utilizando IC, estatística, pesquisa operacional e modelagem numérica. [Hauser-Davis et al. 2010] em seu trabalho teve o intuito de aplicar técnicas de análise discriminante (AD) e rede neural artificial (RNA) para fazer a classificação de três espécies de peixes em diferentes locais no estado do Rio de Janeiro, Brasil. A RNA apresentou uma taxa de acurácia superior a AD nesse estudo.

[Liu and Chen 2012] consideraram a temperatura da água como um fator dominante no controle da estratificação, qualidade da água e de lagos ecológicos, devido muitos processos biológicos serem dependentes da temperatura. Um modelo tridimensional (MT) e RNA é usado para os estudos de simulação da temperatura da água. Os resultados mostram que o desempenho de predição da temperatura da água no MT tem menor taxa de erro do que o modelo de RNA.

[Cavalcante et al. 2013] desenvolveu um estudo comparativo utilizando técnicas de IC e Estatística para classificar os parâmetros do índice de qualidade da água no período chuvoso e seco do reservatório da UHE de Tucuruí e Coracy Nunes. Os resultados mostraram que uma rede RNA teve maior acurácia em relação a AD com valor de 100% e 91% respectivamente. [Bertholdo et al. 2013] aplicou técnicas de IC na clusterização no suporte aos sistemas de gestão ambiental para descobrir regiões hidrográficas homogêneas quanto às suas características físicas, químicas e ecotoxicológicas através das análises de qualidade de água de alguns dos principais rios do estado de São Paulo, realizadas entre 2005 e 2011. Os resultados apresentam que a metodologia desenvolvida contribui para um melhor conhecimento dos corpos d'água, permitindo a redução da quantidade de pontos a serem analisados em programas de monitoramento.

[Borges Pedro et al. 2014] Analisou a influência dos CHs sobre os parâmetros físico-químico da água na parte médio do rio Solimões entre 2004 a 2011 em diversos corpos d'água utilizando estatística descritiva e agrupamento. Os resultados apresentam que a variação do nível da água durante os CHs avaliados ao longo dos sete anos de monitoramento influenciou a qualidade da água nos lagos da Reserva de Desenvolvimento Sustentável Mamirauá e nos rios Solimões e Tefé. As variações mais acentuadas, observadas nos parâmetros transparência, oxigênio dissolvido e condutividade que ocorreram entre os períodos de seca.

Com base nesse contexto, o objetivo deste trabalho consiste em avaliar de forma sistemática algoritmos de IC para classificação de CHs no reservatório da UHE de Tucuruí na região amazônica a partir de PFQM da água, a fim de servir como estimativa para anos posteriores na administração dos recursos hídricos para esses tipos de corpos d'água. Em primeiro momento foi realizada um análise exploratória dos PFQM da água para definir quais eram os mais relevantes a partir das amostras observadas. Após a análise, algoritmos de IC tais como RNA, máquinas de vetores de suporte (SVM-*Support Vector Machine*) com núcleo radial e polinominal, k-vizinhos mais próximos (KNN-K-*Nearest Neighbors*) e random forest (RF) foram avaliados no processo de classificação de CHs.

O artigo está organizado da seguinte forma: na Seção 2 são apresentado a área de estudo, o modelo adotado para o domínio do problema, métodos e recursos utilizados. A Seção 3 descreve as técnicas de IC utilizadas na classificação dos CHs e a ferramenta empregada no experimento. Na Seção 4 é exposto os resultados do estudo. E por fim na Seção 5, na conclusão, é exposto as considerações finais referentes a este trabalho.

2. Métodos e recursos

A metodologia adotada para avaliação dos classificadores é apresentada na Figura 1. A primeira etapa consiste na análise exploratória dos dados. Entre as técnicas mais utilizadas destaca-se a análise fatorial (AF). A partir desta técnica foi possível extrair os atributos mais significativos e excluir os que não tem relevância na classificação dos CHs.

A AF utilizou KMO (*Kaiser-Meyer-Olkin*), que é índice usado para avaliar a adequação da técnica. O teste varia entre 0 e 1, e valores maior ou igual a 0,50 são aceitáveis. Outro mecanismo da AF foi o teste de esfericidade de Bartlett, que é um teste que examina a hipótese das variáveis correlacionadas na população, sendo que a significância para o teste não deve ultrapassar 0,05. E as comunalidades, que explica a porção da variância que um atributo compartilha com todas as outras variáveis consideradas.

Nesse teste valores igual ou acima de 0,50 são aceitáveis [Coletti et al. 2010].

As etapas seguintes tratam efetivamente da avaliação sistemática dos classificadores adotados para classificação dos CHs.

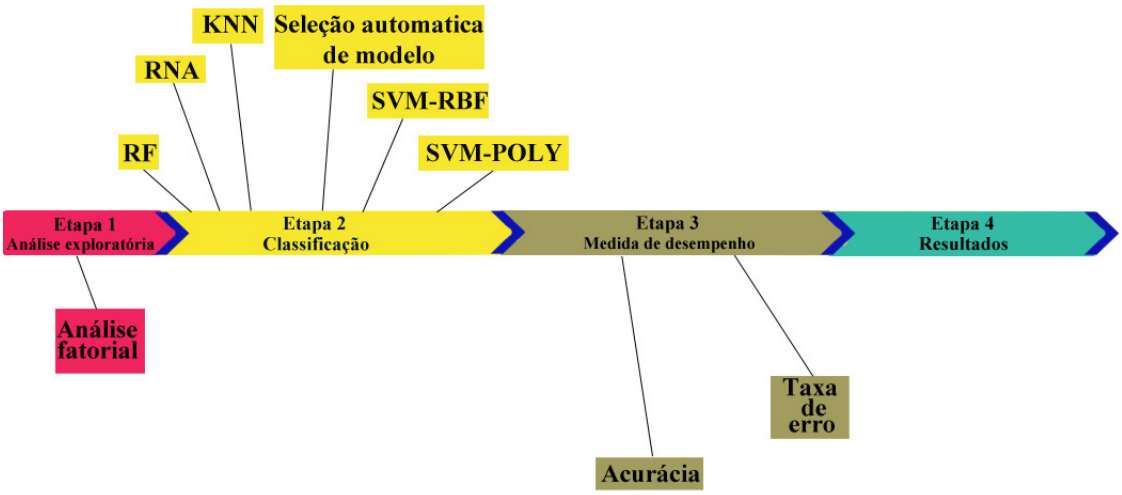

Figura 1. Etapas realizadas para avaliação dos algoritmos de IC utilizados para classificação dos CHs.

2.1. Área de estudo

A área de estudo da pesquisa foi UHE de Tucuruí localizada no Rio Tocantins, no estado do Pará - Brasil, sendo que as coletas e análises foram realizadas no período de janeiro de 2009 a dezembro de 2012. Neste período foram realizadas quatro campanhas em cada ano, considerando os CHs da região e consequentemente o nível de água à montante: seco, enchendo, cheio e vazante em nove estações de amostragem. Estas foram escolhidas de modo a se ter uma amostragem representativa de todo reservatório, dispondo de informações adequadas sobre o reservatório com precisão espacial suficiente [Eletrobras-Eletronorte 2008].

Os PFQM analisados nas águas do reservatório com suas unidades de medidas foram: Secchi (m), Temperatura (°C), pH, OD (mg O2/L), Condutividade (μs cm^{-1}), Ferro (mg L^{-1}), Ca (mg L^{-1}), Mg (mg L^{-1}), K (mg L^{-1}), Na (mg L^{-1}), NH4 (μg L^{-1}), NO3(mg L^{-1}), total PO4 (mg L^{-1}), PO4 (mg L^{-1}), STS(mg L^{-1}), Turbidez (NTU), Clorofila-a (mg L^{-1}). A Figura 2 mostra os sítios amostrais utilizados nas coletas: Caraipé 1(C1), Caraipé 2(C2), Breu Branco(MBB), Ipixuna (MIP), Jacundá Velho (JV), Lontra (ML), Pucuruí (MP), Montante 3 (M3) e Belauto (BE).

2.2. Modelo de classificação adotado

A classificação foi realizada em base de dados fornecida pela Eletrobras-Eletronorte contendo um total de 423 registros de análise de água de nove sítios amostrais do reservatório da UHE de Tucuruí no período de 2009 a 2012. Foram consideradas 17 variáveis de entrada para os classificadores mapeados a partir dos PFQM, objetivando a classificação dos CHs que afetam o nível de água. A Figura 3 apresenta o modelo de classificação dos CHs.

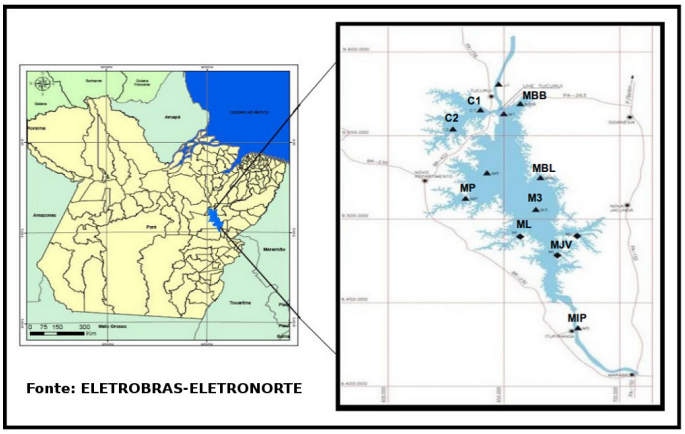

Figura 2. Distribuição das estações de coleta à montante do reservatório.

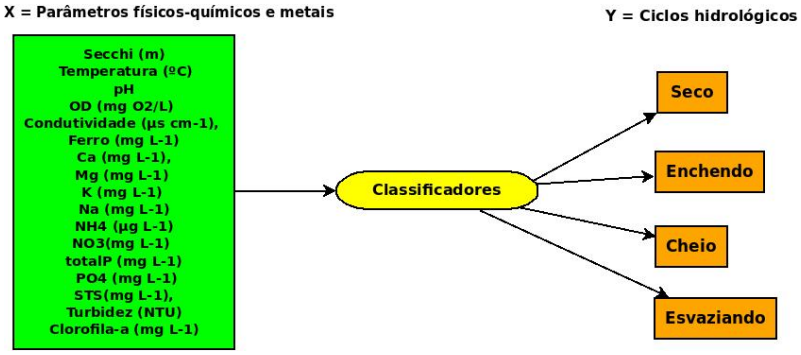

Figura 3. Modelo de classificação dos CHs.

3. Classificadores

Um *classificador* F é um mapeamento $\mathcal{F} : \mathbb{R}^K \rightarrow \{1 \ldots, Y\}$, onde K é a dimensão do vetor de entrada $\mathbf{z} \in \mathbb{R}^K$ e o *rótulo* $y \in \{1, \ldots, Y\}$ é a classe. Quando se treina um classificador usando aprendizado supervisionado [Hastie et al. 2001], é dado um *conjunto de treinamento* $T = \{(\mathbf{z}_1, y_1), ..., (\mathbf{z}_V, y_V)\}$ contendo V *exemplos* de $(\mathbf{z}, y)$.

Um *conjunto de teste* $\{(\mathbf{z}_1, y_1), ..., (\mathbf{z}_R, y_R)\}$ contendo R exemplos e disjunto do conjunto de treino pode ser usado para calcular a taxa de erro de classificação

$$E_f = \frac{1}{R} \sum_{r=1}^{R} \mathcal{I}(F(\mathbf{z}_r) \neq y_r), \tag{1}$$

onde $\mathcal{I}$ é a função indicador, que é um (1) caso o argumento seja verdadeiro e zero (0) caso contrário. O erro E_f é uma estimativa da capacidade de generalização do classificador [Witten and Frank 2005].

Para classificação este trabalho adota os algoritmos disponíveis no software WEKA (*Waikato Environment for Knowledge Analysis*) [1]. Assim, a próxima seção discutirá brevemente este software e as seções seguintes listarão os principais classificadores utiliza-

[1]Software de domínio público desenvolvido na Universidade de Waikato na Nova Zelândia. Disponível em http://www.cs.waikato.ac.nz/ml/weka/

dos. Como o objetivo deste trabalho não é discutir cada um desses classificadores em profundidade, busca-se prioritariamente ilustrar como os mesmos são usados no WEKA.

3.1. Weka

O pacote WEKA é reconhecido pela comunidade científica como um sistema de referência em aprendizado de máquina e mineração de dados [Witten and Frank 2005]. O sistema é formado por um conjunto de implementações de algoritmos de diversas técnicas de aprendizado de máquina, e foi implementado na linguagem de programação Java, tornando-o acessível nas principais plataformas computacionais.

O WEKA inclui algoritmos de regressão, classificação, agrupamento, regras de associação e seleção de parâmetros (atributos). Atualmente está na versão 3.6.10, e antes de utilizá-lo os dados devem ser convertidos para um dos formatos de arquivo suportados pelo WEKA. Neste trabalho o formato adotado é próprio do WEKA denominado de arff (*Attribute-Relation File Format*). O mesmo possui diversos algoritmos de IC e este trabalho usa apenas alguns dos principais, tais como RNA multicamadas treinadas com algoritmo backpropagation, SVM, RF e KNN. [Haykin 2001, Witten and Frank 2005].

Nos classificadores, frequentemente o melhor desempenho de um algoritmo de IC em relação a uma base de dados específica só pode ser alcançada ajustando-se exaustivamente os parâmetros do algoritmo. Esta tarefa é chamada de seleção automática do modelo e corresponde, por exemplo, a escolher os parâmetros tais como o número de neurônios na camada escondida de uma RNA. Uma estratégia popular para seleção automática do modelo e que foi adotada neste trabalho é a validação-cruzada(*cross-validation*) considerando 10-*folds* implementada na classe CVParameterSelection do WEKA [Witten and Frank 2005].

3.2. Redes neurais artificiais

As RNA são sistemas paralelamente distribuídos compostos por unidades simples de processamento denominados de neurônios artificiais que calculam uma certa função matemática (usualmente não-linear). Tais unidades são dispostas em uma ou mais camadas e interligadas pelos chamados pesos sinápticos. O comportamento inteligente de uma RNA vem das interações entre as unidades de processamento da rede.

O algoritmo utilizado neste trabalho para treinamento da RNA MLP foi o backpropagation também chamado de regra delta generalizada [Haykin 2001]. Neste trabalho foi utilizada a classe *MultlayerPerceptron* que é o padrão do WEKA e os principais parâmetros dessa implementação são: (-L) - corresponde à taxa de aprendizado utilizada pelo algoritmo *backpropagation* e este valor deve ser entre 0 e 1 (Padrão é 0,3); (-M) - taxa de momento para o algoritmo *backpropagation* e este valor deve ser entre 0 e 1 (Padrão é 0,2); (-N) - este parâmetro corresponde ao número de épocas para treinamento da rede e o padrão é 500; (-H) - corresponde à quantidade de camadas ocultas que podem ser criadas na rede.

3.3. Máquinas de vetores de suporte

A SVM constitui uma técnica de IC fundamentada pela teoria de aprendizado estatístico [Vapnik 1995]. O objetivo desse classificador consiste em encontrar um hiperplano com uma máxima margem de separação em um espaço de características. Um hiperplano tem

como função ser uma superfície de decisão de tal forma que a margem de separação entre exemplos de uma classe e outra seja máxima [Haykin 2001].

No WEKA o classificador SVM é implementado na forma de um módulo extra. Este classificador implementa uma versão da LibSVM otimizada para uso com o WEKA [EL-Manzalawy and Honavar 2005]. Para lidar com o problema de multi-classes, as SVMs são organizadas na LibSVM no esquema *one-versus-one* (ou *all-pairs*) [Rifkin and Klautau 2004]. No escopo desse trabalho, as funções *kernel* utilizadas são a RBF e Polinomial. Assim, os principais parâmetros utilizados foram o C (C >0), parâmetro de penalidade do termo de erro, e o G, largura das funções *kernel*.

3.4. Random forest

Quando se aborda temas em IC sobre RF, remete-se a árvores de decisão. A proposta é que dividindo o espaço de instâncias em subespaços, os mesmos são ajustados em diferentes modelos, sendo formalmente estruturado em um grafo acíclico em que cada nó, ou é um nó de divisão com dois ou mais sucessores dotado de um teste condicional, ou um nó folha rotulado com uma nova classe [Zhou et al. 2014]. No WEKA o RF tem inúmeros parâmetros, sendo os principais são: (*-depth*) - responsável pela profundidade da árvore; (K) - número de recursos em cada iteração; (I)- número de árvores do modelo.

3.5. K-nearest neighboor

Para [Haykin 2001] o classificador KNN utiliza os próprios dados de treinamento como modelo de classificação, isto é, para cada novo padrão que se quer classificar, usa os dados do treinamento para verificar quais são os exemplos nessa base de dados que são "mais próximos" do padrão em análise, ou seja, baseado nas distâncias. A cada novo padrão a ser classificado faz-se uma varredura nos dados de treinamento, o que provoca um grande esforço computacional. O KNN no WEKA é implementado na classe *IBK* e seus principais parâmetros são: (-N) que é igual ao número de centros (ou K) e (-S) que gera aleatoriamente os centros.

4. Resultados experimentais

Os resultados da AF teve como objetivo diminuir o número de variáveis. Neste estudo, através do KMO maior que 0,50, foi avaliado adequação da AF para a proposta do estudo. O teste de esfericidade de Bartlett menor que 0,05 para verificar a correlação entre as variáveis. E as comunalidades acima de 0,50 que explica a porção da variância compartilhada entre todas variáveis consideradas. A Tabela 1 apresenta o resultados da AF.

Conforme a Tabela 1, o resultado do KMO ficou em 0,763, acima de 0,5 e validando que a AF está adequada para proposta do estudo. O teste de esfericidade de Bartlett apresentou valor de significância de 0,000, sendo menor que 0,50, ajustando a correlação entre as variáveis. E os valores de extração dos fatores estão todos com suas comunalidades também acima de 0,50. Assim, a AF selecionou as variáveis com maior relevância que represente modelo esperado. E na seleção automática do modelo foi possível a escolha dos melhores valores para os parâmetros dos classificadores. A Tabela 2 apresenta o *grid* de parâmetros adotado para o procedimento de seleção automática do modelo.

Após vários testes a arquitetura mais robusta para RNA foi com apenas uma camada escondida com 43 neurônios fixados em H, taxa de aprendizagem L de 0.3 e taxa

Tabela 1. Resultado da análise fatorial.

KMO	=	0,763
Teste de esfericidade de Bartlett	sig	0,000
=	Comunalidades	=
Variáveis	Inicial	Extração
Transp	1	0,772
Temp	1	0,541
OD	1	0,645
pH	1	0,688
Cond	1	0,660
FeTotal	1	0,725
Ca	1	0,689
Mg	1	0,726
Na	1	0,814
K	1	0,763
NH_4	1	0,728
NO_3	1	0,686
PO_4	1	0,622
PTOTAL	1	0,838
STS	1	0,713
PigTOTAL	1	0,703
Turb	1	0,761

Tabela 2. *Grid* de parâmetros adotado no procedimento de seleção automática do modelo.

Classificador	Parâmetros	Valores do *Grid*	Melhor Valor
RNA	H	11, 43, 75, 107, 139 e 171	43
	L	0,1, 0,3, 0.5, 0,7, e 0,9	0,3
	M	0,1, 0,3 e 0,9	0,3
RF	I	100, 200, 300, ..., e 1000	100
SVM-RBF	G	0,1, 1, 10 e 100	10
	C	0,1, 1, 10 e 100	100
SVM-POLY	G	0,1, 1, 10 e 100	1
	C	0,1, 1, 10 e 100	10
KNN	K	1, 3, 5, 7, 9, 11, 13 e 15	1

de momentum M de 0,3, sendo que o número de épocas fixado em 2000. Para os classificadores SVMs, os parâmetros otimizados foram gama G e penalidade do erro C. Depois de vários experimentos o melhor desempenho para SVM-RBF foi 10 em G e 100 em C e para o classificador SVM-POLY foi 1 em G e 10 para C. Para o classificador Random Forest a seleção de modelo levou em consideração o principal parâmetro deste classificador o número de árvores a ser gerada I. Para o classificador KNN, K é o número de vizinhos com o valor 1.

A Tabela 3 apresenta os desempenhos dos classificadores utilizados levando em consideração a taxa classificação incorreta e a acurácia ($A_c = 1 - E_f$) como medidas de desempenho.

Tabela 3. Taxa de erro dos classificadores utilizados.

Classificador	Taxa de erro (E_f)%	Acurácia (A_c)%
RNA	15,5	84,5
KNN	9,9	90,1
SVM-RBF	10,9	89,1
SVM-POLY	14,5	85,5
RF	7,8	92,2

De acordo com os resultados apresentados na Tabela 3, pode-se observar que o classificador Random Forest foi o que teve o melhor desempenho apresentando uma taxa

de erro 7,8% sendo que os classificadores RNA e SVM-POLY foram os que apresentaram um pior desempenho, apesar de seus resultados serem considerados relativamente satisfatórios dado a complexidade em classificar os diferentes CHs.

5. Conclusão

O presente trabalho avaliou de forma sistemática diferentes classificadores na predição dos CHs levando em consideração a alteração dos PFQM da água. As métricas de desempenho adotadas de taxa de erro e acurácia foram utilizadas para validar os resultados obtidos. Entre as técnicas utilizadas, o classificador Random Forest foi o que apresentou melhor desempenho. Vale ressaltar que a grande variação sazonal climática da região amazônica pode ter interferido na obtenção de resultados mais precisos. Essa precisão também é muitas vezes afetada, pois as concentrações dos PFQM da água não são homogêneos nas estações de amostragem.

Portanto, a aplicação de técnicas de IC na classificação de CHs tem uma contribuição construtiva na administração dos recursos hídricos em reservatórios de água na amazônia. Possibilitando pesquisas futuras que possam utilizar outras técnicas computacionais, além das empregadas no estudo, para auxíliar de forma mais efetiva o gerenciamentos dos recursos hídricos nestes locais.

Agradecimentos

Os autores agradecem a Eletrobras-Eletronorte por ter fornecido as amostras de análise de água dos sítios amostrais da UHE de Tucuruí que possibilitou a realização da pesquisa.

Referências

Bertholdo, L., Júnior, L. C., de Aragão Umbuzeiro, G., and da Silva, C. G. (2013). Mineração de dados de qualidade de água para agrupamento de pontos de amostragem usados no monitoramento de recursos hídricos. *WCAMA - CSBC*, 1:1036–1046.

Bittencourt, M. M. and Amadio, S. A. (2007). Proposta para identificação rápida dos períodos hidrológicos em áreas de várzea do rio solimões-amazonas nas proximidades de manaus. *Acta Amazonica*, 37(2):303–308.

Borges Pedro, J. P., Rosinski Lima Gomes, M. C., de Jesus Trindade, M. E., Pedrociane Cavalcante, D., Alves de Oliveira, J., Pucci Hercos, A., Zucchi, N., Barbosa de Lima, C., Aquino Pereira, S., and Lima de Queiroz, H. (2014). Influence of the hydrological cycle on physical and chemical variables of water bodies in the várzea areas of the middle solimões river region (amazonas, brazil). *UAKARI*, 9(2):75–90.

Cavalcante, Y., Hauser-Davis, R., Saraiva, A., Brandão, I., Oliveira, T., and Silveira, A. (2013). Metal and physico-chemical variations at a hydroelectric reservoir analyzed by multivariate analyses and artificial neural networks: Environmental management and policy/decision-making tools. *Science of the Total Environment*, 442:509–514.

Coletti, C., Testezlaf, R., Ribeiro, T. A., de Souza, R. T., and Pereira, D. d. A. (2010). Water quality index using multivariate factorial analysis. *Revista Brasileira de Engenharia Agrícola e Ambiental*, 14(5):517–522.

EL-Manzalawy, Y. and Honavar, V. (2005). *WLSVM: Integrating LibSVM into Weka Environment*. Software available at `http://www.cs.iastate.edu/~yasser/wlsvm`.

Eletrobras-Eletronorte (2008). *Manual do sistema de gestão ambiental - UHE Tucuruí*. Eletrobras-Eletronorte, Tocantins.

Galvão, C. d. O. (1999). *Sistemas inteligentes: Aplicações a recursos hídricos e ciências ambientais*. UFRGS: ABRH.

Gastaldini, M., SEFFRIN, G., and Paz, M. (2002). Diagnóstico atual e previsão futura da qualidade das águas do rio ibicuí utilizando o modelo qual2e. *Engenharia sanitária e ambiental*, 7(3):129–138.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The elements of statistical learning*. Springer Verlag.

Hauser-Davis, R., Oliveira, T., Silveira, A., Silva, T., and Ziolli, R. (2010). Case study: Comparing the use of nonlinear discriminating analysis and artificial neural networks in the classification of three fish species: acaras (geophagus brasiliensis), tilapias (tilapia rendalli) and mullets (mugil liza). *Ecological Informatics*, 5(6):474–478.

Haykin, S. (2001). *Redes Neurais: Principios e Prática. 2. Ed.* Porto Alegre: Bookman.

Liu, W. and Chen, W. G. (2012). Prediction of water temperature in a subtropical subalpine lake using an artificial neural network and three-dimensional circulation models. *Computers & Geoscience*, 45:13–25.

Rifkin, R. and Klautau, A. (2004). In defense of one-vs-all classification. *J. Machine Learning Research*, 5:101–141.

Singh, K. P., Basant, A., Malik, A., and Jain, G. (2009). Artificial neural network modeling of the river water quality: a case study. *Ecological Modelling*, 220(6):888–895.

Tundisi, J. G. (1999). *Limnologia no século XXI: perspectivas e desafios*. Instituto Internacional de Ecologia.

Vapnik, V. (1995). *The nature of statistical learning theory*. Springer Verlag.

Vrana, B., Allan, I. J., Greenwood, R., Mills, G. A., Dominiak, E., Svensson, K., Knutsson, J., and Morrison, G. (2005). Passive sampling techniques for monitoring pollutants in water. *TrAC Trends in Analytical Chemistry*, 24(10):845–868.

Witten, I. H. and Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Zhou, Q., Zhou, H., Zhou, Q., Yang, F., and Luo, L. (2014). Structure damage detection based on random forest recursive feature elimination. *Mechanical Systems and Signal Processing*.