

11º CONGRESSO BRASILEIRO DE PESQUISA E DESENVOLVIMENTO EM PETRÓLEO E GÁS



TÍTULO DO TRABALHO:

CLASSIFICAÇÃO DE SEDIMENTOS DA MARGEM EQUATORIAL BRASILEIRA UTILIZANDO ALGORITMOS DE APRENDIZADO DE MÁQUINA

AUTORES:

Luiz Eugênio Hoffmann Lopes¹, Allan Kardec Duailibe Barros Filho¹, Cleverson Guizan Silva², Diego de Oliveira Dantas¹, João Regis dos Santos Filho², Marta de Oliveira Barreiros¹.

INSTITUIÇÃO:

¹Departamento de Engenharia Elétrica, Laboratório de Processamento de Informação Biológica (PIB), Universidade Federal do Maranhão (UFMA).

²Departamento de Geologia, LAGEMAR, Universidade Federal Fluminense (UFF).

CLASSIFICAÇÃO DE SEDIMENTOS DA MARGEM EQUATORIAL BRASILEIRA UTILIZANDO ALGORITMOS DE APRENDIZADO DE MÁQUINA

Resumo

Recentemente, importantes discussões a respeito da ocorrência de extensos recifes carbonáticos e do potencial de descobertas de hidrocarbonetos nas bacias do Pará-Maranhão e Foz do Amazonas, elevaram a importância do estudo da composição faciológica da plataforma continental. No entanto, amostrar sedimentos marinhos é uma operação de alto custo e muitas vezes demorada a depender da distância da costa e da profundidade da amostragem, sendo necessário o deslocamento da equipe e dos equipamentos através de embarcações de pesquisa especializadas. Tendo isto em vista, o objetivo principal deste trabalho foi integrando informações sedimentológicas de amostras de fundo e atributos sísmicos a partir de algoritmos de aprendizado de máquina para assim otimizar o mapeamento das fácies na plataforma continental da Margem Equatorial Brasileira. Os dados de sedimentos foram adquiridos do Banco Nacional de Dados Oceanográficos (BNDO), sendo classificados em Lama, Areia e Cascalho para o tamanho do grão e em Litoclástico (<30%), Litobioclástico (<50%), Biolitoclástico (<70%) e Bioclástico (>70%) para o teor de carbonato de cálcio. Os atributos sísmicos utilizados foram Coerência, Amplitude, Intensidade, Fase Instantânea e Frequência Instantânea. Foram desenvolvidos cinco algoritmos de aprendizado de máquina com a finalidade de relacionar os tipos de sedimento e teor de carbonatos aos atributos sísmicos. Os resultados de cada algoritmo são: *Support Vector Machine*: Accuracy 80,31%, Precision 81,21%, Recall 80,31% e F1_Score 80,01%. *K-Nearest Neighbors*: Accuracy 77,07%, Precision 80,55%, Recall 77,07% e F1_Score 75,94%. *Decision Trees*: Accuracy 85%, Precision 85,78%, Recall 85% e F1_Score 84,85%. *Multilayer Perceptron*: Accuracy 82,74%, Precision 82,74%, Recall 82,74% e F1_Score 82,49%. *Adaptive Boosting*: Accuracy 70,47%, Precision 72,25%, Recall 70,47% e F1_Score 69,75%. O melhor resultado obtido foi do algoritmo *Decision Trees* com 85% de acurácia. Com esses resultados preliminares, observa-se que é possível criar um modelo de aprendizado de máquina para classificar automaticamente os sedimentos marinhos, viabilizando e otimizando pesquisas futuras sobre as fácies carbonáticas e recifes mesofóticos na Margem Equatorial Brasileira.

Palavras-chave: Aprendizado de Máquina, Classificação, Margem Equatorial Brasileira, Sedimentos Marinhos.

Abstract

Recently, important discussions about the occurrence of extensive carbonate reefs and the potential for hydrocarbon discoveries in the Pará-Maranhão and Foz do Amazonas basins have raised the importance of studying the faciological composition of the continental shelf. However, sampling marine sediments is a high-cost and often time-consuming operation, depending on the distance from the coast and the depth of sampling, requiring the displacement of staff and equipment through specialized research vessels. With this in mind, the main objective of this work was to integrate sedimentological information from bottom samples and seismic attributes from machine learning algorithms to optimize the mapping of facies on the continental shelf of the Brazilian Equatorial Margin. Sediment data were acquired from the National Oceanographic Data Bank (BNDO), classified as Mud, Sand and Gravel for grain size and Lithoclastic (<30%), Lithobioclastic (<50%), Biolithoclastic (<70%) and Bioclastic (>70%) for calcium carbonate content. The seismic attributes used were Coherence, Amplitude, Intensity, Instantaneous Phase and Instantaneous Frequency. Five machine learning

algorithms were developed in order to relate sediment types and carbonate content to seismic attributes. The results of each algorithm are: Support Vector Machine: Accuracy 80.31%, Precision 81.21%, Recall 80.31% and F1_Score 80.01%. K-Nearest Neighbors: Accuracy 77.07%, Precision 80.55%, Recall 77.07% and F1_Score 75.94%. Decision Trees: Accuracy 85%, Precision 85.78%, Recall 85% and F1_Score 84.85%. Multilayer Perceptron: Accuracy 82.74%, Precision 82.74%, Recall 82.74% and F1_Score 82.49%. Adaptive Boosting: Accuracy 70.47%, Precision 72.25%, Recall 70.47% and F1_Score 69.75%. The best result was obtained from the Decision Trees algorithm with 85% accuracy. With these preliminary results, it is possible to create a machine learning model to automatically classify marine sediments, enabling and optimizing future research on carbonate facies and mesophotic reefs in the Brazilian Equatorial Margin.

Keywords: Classification, Brazilian Equatorial Margin, Machine Learning, Marine Sediments.

Introdução

A crescente demanda energética tem elevado a importância da Margem Equatorial Brasileira como uma nova fronteira para a extração de recursos não-renováveis como o petróleo e o gás natural, uma vez que a mesma tem semelhanças geológicas com a costa ocidental da África, que já é reconhecida mundialmente por suas reservas petrolíferas. Como as margens Brasileira e Africana são análogas, é possível ter as mesmas condições petrolíferas no Brasil (FILHO et al, 2020), conforme confirmadas pelas recentes descobertas de reservatórios nas bacias de Barreirinhas e Pará-Maranhão (PELLEGRINI e RIBEIRO, 2018). Ao todo, cerca de 18 campos produtores de petróleo já estão consolidados na Margem Equatorial, principalmente nas Bacias Potiguar e Ceará (ANP, 2020).

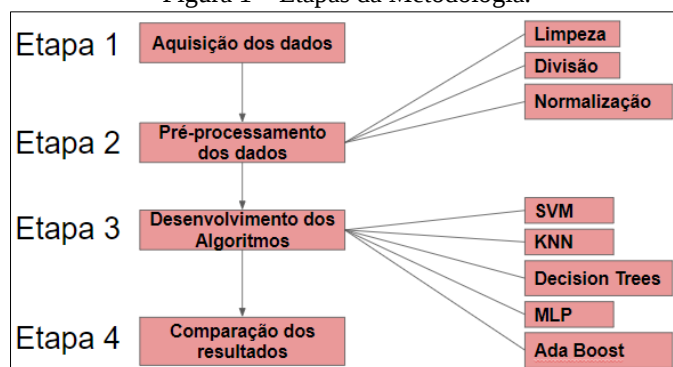
Além disso, os aproximadamente 9.650 km de litoral apresentam condições favoráveis para a implementação de fazendas eólicas offshore no Brasil, cujas políticas de regulação voltadas ao licenciamento, monitoramento e proteção ambiental ainda estão em discussão (GONZÁLEZ *et al.*, 2020). Nesse contexto, as medidas regulatórias, o planejamento estratégico e o uso do espaço marítimo passam pela necessidade de se refinar o reconhecimento do leito marinho onde serão instalados os aerogeradores. Apenas na Margem Equatorial estão em análise até o momento 31 áreas de Complexos Eólicos Offshore sobre a plataforma continental (IBAMA, 2022).

Como o Brasil tem uma extensa região costeira, e uma margem continental adjacente que por vezes ultrapassa as 200 milhas náuticas da sua Zona Econômica Exclusiva (ZEE), a realização da amostragem extensiva dos sedimentos marinhos não é eficiente ou operacionalmente viável para o ideal reconhecimento das fácies sedimentares da plataforma continental. Este trabalho utilizou dados públicos de sedimento e sísmica multicanal 2D disponíveis na região da Margem Equatorial Brasileira para prever qual é a sua classe de sedimento com o auxílio de aprendizado de máquina.

Metodologia

Este trabalho foi dividido em 4 etapas: Aquisição dos dados, Pré-processamento dos dados, Desenvolvimento dos algoritmos e Comparação dos resultados. A figura 1 abaixo, mostra a sequência de etapas realizadas.

Figura 1 – Etapas da Metodologia.



Fonte: Autor.

Aquisição dos dados

Ao todo, foram reunidos 5.515 dados de sedimentos superficiais provenientes do Banco Nacional de Dados Oceanográficos (BNDO) e do Banco de Dados da Indústria do Petróleo (BAMPETRO), apresentando duas informações básicas: natureza do fundo oceânico (feita por descrição visual) ou através de dados sedimentológicos processados em laboratório. Os teores de tamanho de grão foram classificados segundo a classificação ternária de SHEPARD (1954) em três classes principais, sendo: Lama, Areia e Cascalho, e o teor de carbonato de cálcio em quatro classes, sendo: Litoclástico ($\text{CaCO}_3 < 30\%$), Litobioclástico ($30\% < \text{CaCO}_3 < 50\%$), Litobioclástico ($50\% < \text{CaCO}_3 < 70\%$) e Litoclástico ($\text{CaCO}_3 > 70\%$) (LARSONNEUR, 1977 adaptado por DIAS, 1996).

Os dados de atributos sísmicos consistem em resultados da interpretação e processamento de linhas públicas de sísmica 2D multicanal extraídos no software Paleoscan™. Ao todo foram utilizadas 384 linhas sísmicas. Os primeiros 500 milissegundos de cada linha sísmica tiveram seu fundo do mar interpretado manualmente para a extração de cinco atributos sísmicos: amplitude, intensidade de reflexão, fase, frequência e coerência.

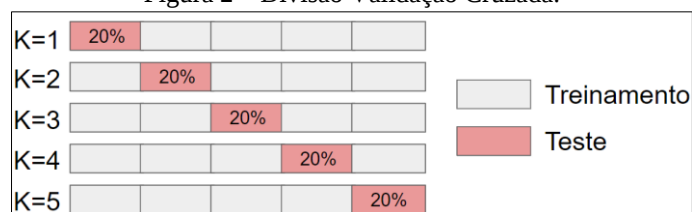
O resultado da etapa 1 foi um arquivo ".csv" com as 5.515 linhas referentes aos pontos de sedimento e 106 colunas dos mais diversos tipos de parâmetros sedimentológicos. Na etapa seguinte, foi realizado o pré-processamento desses dados para serem utilizados na entrada dos algoritmos.

Pré-processamento dos dados

No pré-processamento dos dados foi feita a limpeza, a divisão e a normalização. Inicialmente são excluídos os dados inconsistentes ou nulos. Depois são separadas apenas as colunas de interesse: Classe_Shepard (Areia, Cascalho, Lama), a latitude e a longitude do sedimento, e os cinco atributos sísmicos associados. O resultado foi um arquivo ".csv" com 8 colunas, 101 linhas de areia, 74 linhas de cascalho e 73 linhas de lama.

A Figura 2 abaixo, mostra como foi realizada a Validação Cruzada dos dados. Se K for igual a 5, significa que os dados são divididos em 5 partes (20% dos dados), ou seja, o algoritmo executa 5 vezes, onde na primeira execução a parte 1 será destinada para teste e o restante para treinamento. Já na segunda execução, a parte 2 dos dados será destinada para teste e o restante para treinamento, assim sucessivamente até $K = 5$.

Figura 2 – Divisão Validação Cruzada.



Fonte: Autor.

A técnica de Validação Cruzada é utilizada para avaliar a capacidade de generalização de um modelo, com ela é possível realizar o teste do algoritmo em todos os dados.

Desenvolvimento dos algoritmos

Neste trabalho foram desenvolvidos cinco algoritmos de aprendizado de máquina, descritos a seguir:

i- *Support Vector Machine* (SVM), ou Máquinas de Vetores de Suporte, constituem uma técnica de Aprendizado de Máquina embasada na teoria de aprendizado estatístico, desenvolvida por VAPNIK (1995). Tem uma série de princípios para obter classificadores com boa generalização, ou seja, boa capacidade de prever corretamente a classe de novos dados do mesmo domínio em que o aprendizado ocorreu (VALDECY, 2013).

O fundamento básico do funcionamento de uma SVM é a construção de um hiperplano (objeto matemático), o qual pode possuir uma dimensão R^n , possibilitando a separação dos dados em classes distintas. Essa separação tem abordagem baseada na Teoria de Aprendizagem Estatística, utilizando a otimização matemática, para promover a busca de um hiperplano ótimo, adotando os mínimos quadrados por meio da minimização estrutural de risco (BARRETO et al, 2019).

ii- *K-Nearest Neighbors* (KNN) é um método de classificação muito popular em mineração de dados, regressão e imputação de valor ausente, por causa de sua implementação simples e desempenho de classificação significativo. Realiza tarefas de classificação calculando primeiro a distância entre a amostra de teste e todas as outras amostras de treinamento para obter seus vizinhos mais próximos e, em seguida, classifica as amostras de teste com rótulos pela regra da maioria nos rótulos de vizinhos mais próximos selecionados (ZHANG et al, 2018).

iii- *Decision Trees*, ou Árvores de decisão, são modelos de processos biológicos e cognitivos, são formas simples e eficazes de representar o conhecimento. Para construção de uma *Decision Trees*, seleciona-se inicialmente um atributo de particionamento e divide o conjunto de exemplos, criando um ramo para cada valor deste atributo. A cada ramo é criado um nodo, e novamente selecionado um novo atributo de particionamento para o subconjunto de exemplos atualizando o nodo e, assim, sucessivamente. Tem como objetivo separar as classes em partições diferentes (GARCIA, 2003).

iv- *Multilayer Perceptron* (MLP) é uma rede neural artificial com uma ou mais camadas ocultas e um número indeterminado de neurônios, é uma derivação do *Perceptron* simples. O número de camadas ocultas em um *Perceptron* multicamadas e o número de neurônios em cada camada pode variar dependendo do problema. Em geral, mais neurônios oferecem maior sensibilidade ao problema a ser resolvido, mas também maior risco de *overfit* (MURTAGH, 1991).

v- *Adaptive Boosting*, também conhecido como *AdaBoost*, pode ser utilizado para aumentar a performance de outros algoritmos de aprendizagem. É adaptável no sentido de que as classificações subsequentes feitas são ajustadas a favor das instâncias classificadas negativamente por classificações anteriores. É sensível ao ruído nos dados e casos isolados. Os principais diferenciais desse algoritmo são: (i) requer baixo valor computacional, e (ii) a margem das fronteiras de decisão entre as classes é mais precisa que outros algoritmos predecessores (CHAVES, 2011).

Comparação dos resultados

As métricas mais utilizadas em aprendizado de máquina para avaliação da qualidade dos classificadores são: *Accuracy*, *Precision*, *Recall* e *F1-Score*. Onde: *Accuracy* – indica uma performance geral do modelo, ou seja, dentre todas as classificações, quantas o modelo classificou corretamente; *Precision* – encontra os Falsos Negativos; *Recall* – encontra os Falsos Positivos; *F1-Score* – é a média harmônica entre *Precision* e *Recall*.

Para realizar o trabalho, foi utilizada a linguagem de programação *python* e a biblioteca *sklearn* no ambiente *Google Colab Pro*, que fornece o seguinte *hardware*: GPU Tesla P100-PCIE-16GB, Intel Xeon @ 2x 2 GHz CPU, 13.021 MiB RAM, 70 GB HD e *Ubuntu 18.04 Linux Kernel 4.19.104*.

Resultados e Discussão

Foram realizados testes com os dados de latitude, longitude do sedimento e os cinco atributos sísmicos: amplitude, intensidade de reflexão, fase, frequência e coerência. Mas os melhores resultados foram encontrados utilizando latitude, longitude e apenas três dos atributos sísmicos: amplitude, fase e frequência. Os melhores resultados obtidos são apresentados nas tabelas a seguir.

Na Tabela 1 abaixo, são apresentados os resultados dos classificadores MLP, KNN e SVM. O MLP tem a seguinte configuração: 50 camadas ocultas, 100 neurônios por camadas, taxa de aprendizado 0,000001, função de ativação RELU e otimização ADAM. O algoritmo KNN tem como configuração o valor de $K = 3$, ou seja, os três vizinhos mais próximos são utilizados para classificação. O SVM tem a seguinte configuração: $C = 1$, kernel = rbf, degree = 3, gamma = scale.

Tabela 1 – Resultados dos Classificadores MLP, KNN e SVM.

K-FOLD	MLP				KNN				SVM			
	Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score
K = 1	75,51%	75,25%	75,51%	74,74%	79,59%	83,67%	79,59%	78,72%	77,55%	80,88%	77,55%	76,80%
K = 2	87,75%	87,52%	87,75%	87,58%	81,63%	86,87%	81,63%	80,77%	87,75%	87,74%	87,75%	87,21%
K = 3	77,55%	78,08%	77,55%	77,50%	71,42%	72,99%	71,42%	68,67%	73,46%	74,39%	73,46%	73,34%
K = 4	85,14%	85,10%	85,14%	85,03%	77,22%	80,56%	77,22%	76,67%	81,18%	81,31%	81,18%	81,09%
K = 5	87,75%	87,77%	87,75%	87,59%	75,51%	78,65%	75,51%	74,84%	81,63%	81,74%	81,63%	81,63%
MÉDIA	82,74%	82,74%	82,74%	82,49%	77,07%	80,55%	77,07%	75,94%	80,31%	81,21%	80,31%	80,01%

Fonte: Autor.

Na Tabela 2 abaixo, são apresentados os resultados dos classificadores *Decision Trees* e *AdaBoost*. Suas implementações foram feitas com as configurações padrões, conforme encontrada na biblioteca *sklearn*.

Tabela 2 – Resultados dos Classificadores *Decision Trees* e *AdaBoost*.

K-FOLD	Decision Trees				AdaBoost			
	Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score
K = 1	91,83%	92,34%	91,83%	91,88%	89,79%	91,29%	89,79%	89,90%
K = 2	85,71%	86,88%	85,71%	85,94%	69,38%	69,22%	69,38%	67,64%
K = 3	75,51%	76,84%	75,51%	74,84%	73,46%	75,32%	73,46%	74,06%
K = 4	82,17%	82,85%	82,17%	81,86%	56,43%	60,75%	56,43%	55,82%
K = 5	89,79%	89,98%	89,79%	89,71%	63,26%	64,65%	63,26%	61,34%
MÉDIA	85,00%	85,78%	85,00%	84,85%	70,47%	72,25%	70,47%	69,75%

Fonte: Autor.

Conclusões

Neste trabalho foram desenvolvidos cinco algoritmos para classificação dos sedimentos marinhos da Margem Equatorial Brasileira: Multilayer Perceptron, K-Nearest Neighbors, Decision Trees, Adaptive Boosting e Support Vector Machine. Os dois melhores resultados encontrados foram no algoritmo MLP com *Accuracy* 82,74%, *Precision* 82,74%, *Recall* 82,74% e *F1_Score* 82,49% e no algoritmo *Decision Trees* com *Accuracy* 85%, *Precision* 85,78%, *Recall* 85% e *F1_Score* 84,85%.

Com esses resultados, observa-se que é possível criar um modelo de aprendizado de máquina para classificar automaticamente os sedimentos marinhos na área de estudo. Espera-se que esses algoritmos possam otimizar o mapeamento de fácies em toda a plataforma continental, subsidiando pesquisas futuras quanto a classificação das fácies sedimentares, auxiliando nas tomadas de decisão quanto ao ordenamento do uso espacial marítimo e na regulação das diversas atividades industriais na Margem Equatorial Brasileira.

Agradecimentos

Agradecemos ao apoio financeiro da Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP) e da Financiadora de Estudos e Projetos (FINEP), por meio do Programa de Recursos Humanos da ANP para o Setor Petróleo e Gás - PRH/ANP (PRH 54.1 - UFMA). Também agradecemos a Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP) e a TGS pelo fornecimento dos dados de sísmica multicanal.

Referências Bibliográficas

ANP. **BDEP: Banco de Dados de Exploração e Produção**. Disponível em: <<http://www.anp.gov.br/publicacoes/folderes/2416-bdep-banco-de-dados-de-exploracao-e-producao>>.

Acesso em: 7 maio. 2022.

BARRETO, L. C.; CARVALHO, F. O.; SANTOS, B. C. V.; SOARES, A. P. M. R.; SILVA, G. D.. Aplicação de Support Vector Machine na Detecção de Falhas em Sensores de uma Coluna Debutanizadora. XIII Congresso Brasileiro de Engenharia Química em Iniciação Científica, **Blucher Chemical Engineering Proceedings**, Volume 1, 2019, Pages 3003-3009, ISSN 2359-1757. Disponível em: <http://dx.doi.org/10.1016/cobecic2019-SOCP31>.

CHAVES, B. B.. **Estudo do algoritmo AdaBoost de aprendizagem de máquina aplicado a sensores e sistemas embarcados**. 2011. Dissertação (Mestrado em Engenharia de Controle e Automação Mecânica) - Escola Politécnica, Universidade de São Paulo, São Paulo, 2011.

FILHO, A. K. D. B.; CARMONA, R. G.; ZALÁN, P. V.. Nota técnica sobre a margem equatorial brasileira. **Um novo “Pré-sal” no arco norte do território brasileiro?**. 2020.

GARCIA, S. C.. **O uso de árvores de decisão na descoberta de conhecimento na área da saúde**. 2003. Dissertação (Mestrado em Ciência da computação) - Universidade Federal do Rio Grande do Sul.

GONZÁLEZ, M. O. A.; SANTISO, A. M.; MELO, D. C. DE; VASCONCELOS, R. M. DE. Regulation for offshore wind power development in Brazil. **Energy Policy**, v. 145, n. June 2019, 2020.

IBAMA. **Mapas de projetos em licenciamento - Complexos Eólicos Offshore**. Disponível em: <<https://www.ibama.gov.br/laf/consultas/mapas-de-projetos-em-licenciamento-complexos-eolicos-offshore>>. Acesso em: 10 mar. 2022.

LARSONNEUR, C.. 1977. La cartographie des dépôts meubles sur le plateau continental français: méthode mise au point et utilisée en Manche. **J. Rech. Ocean**. 2, 34–39.

MURTAGH, F.. Multilayer perceptrons for classification and regression. **Neurocomputing**. Volume 2. Issues 5–6. 1991. Pages 183-197. ISSN 0925-2312. Disponível em: [https://doi.org/10.1016/0925-2312\(91\)90023-5](https://doi.org/10.1016/0925-2312(91)90023-5).

PELLEGRINI, B. S.; RIBEIRO, H. J. P. S.. Exploratory plays of Pará-Maranhão and Barreirinhas basins in deep and ultra-deep waters, Brazilian Equatorial Margin. **Brazilian Journal of Geology**, v. 48, n. 3, p. 485–502, 2018.

SHEPARD, F. P.. 1954. Nomenclature Based on Sand-silt-clay Ratios. **SEPM J. Sediment. Res.** doi:10.1306/d4269774-2b26-11d7-8648000102c1865d.

VALDECY, V. J.. **Classificação de dados usando técnicas de datamining e aprendizado de máquina**. 2013. Trabalho de Conclusão de Curso (Bacharelado em Ciência da Computação) - Universidade Federal do Maranhão.

VAPNIK, V. N., 1995. The nature of statistical learning theory. **Springer-Verlag New York, Inc.**, New York, NY, USA. ISBN 0-387-94559-8.

ZHANG, S.; LI, X.; ZONG, M.; ZHU, X.; WANG, R.. Efficient kNN Classification With Different Numbers of Nearest Neighbors. 2018. **IEEE Transactions on Neural Networks and Learning Systems**, vol. 29, no. 5, pp. 1774-1785. Disponível em: <https://doi.org/10.1109/TNNLS.2017.2673241>.