8º CONGRESSO BRASILEIRO DE PESQUISA E DESENVOLVIMENTO EM PETRÓLEO E GÁS

TÍTULO DO TRABALHO:

Comparação de Técnicas de Inteligência Computacional para a Classificação de Petrofácies Sedimentares

AUTORES:

Camila Martins Saporetti[1], Leonardo Goliatt da Fonseca[1], Leonardo da Costa Oliveira[2], Egberto Pereira[3]

INSTITUIÇÃO:

1 Universidade Federal de Juiz de Fora

2 Petrobras

3 Universidade Estadual do Rio de Janeiro

COMPARAÇÃO DE TÉCNICAS DE INTELIGÊNCIA COMPUTACIONAL PARA A CLASSIFICAÇÃO DE PETROFÁCIES SEDIMENTARES

Abstract

This work aimed to evaluate a methodology to measure different techniques based on computational intelligence to predict the classification of new thin-sections, which are removed by the same sedimentary basin. The data from boreholes drilled in the Parana Basin was processed in order to validate the proposed methodology. The presented methodology, besides evaluating the performance of various techniques of computational intelligence, is an alternative to assist the geologist/expert in the petrofacies determination and characterization, helping to reduce the manual effort involved in individualization process.

Introdução

Os métodos para a previsão e a determinação da distribuição de heterogeneidades em reservatórios são uma ferramenta importante para exploração e otimização da produção de campos de petróleo [1]. Para ser julgada como um bom reservatório, uma rocha deve possuir: (i) uma extensão considerável, (ii) boa porosidade, (iii) uma adequada permeabilidade e (iv) um eficiente fator de recuperação de hidrocarbonetos, entre outros. Tais características, denominadas petrofísicas, estão ligadas ao percurso deposicional das unidades deposicionais, em especial às condições de sedimentação e a história diagenética, sendo as mesmas de extrema importância para definição da qualidade do reservatório. As heterogeneidades são caracterizadas por meio de diversas petrofácies sedimentares, um conjunto de características petrográficas que individualizam um grupo de rochas.

A definição de petrofácies sedimentares pode ser realizada através da contagem em microscópio petrográfico de luz transmitida, a fim de avaliar as heterogeneidades presentes no reservatório estudado. Este processo usualmente é extenso, pois envolve o processo de amostragem, geração dos dados e posterior interpretações dos mesmos. Contudo nem toda informação é aproveitada, devido à grande quantidade de dados produzidos. Consequentemente, torna-se interessante a mudança do uso de métodos manuais para análises automáticas por meio de ferramentas computacionais. Nesse contexto, técnicas de Inteligência Computacional aparecem como um mecanismo útil para auxiliar na classificação de petrofácies. Técnicas oriundas dessa área têm sido usadas para auxiliar na tomada de decisões de especialistas em diversos problemas de Geociências [2,3, 14,15].

A Inteligência Computacional, procura através de técnicas inspiradas na Natureza, o desenvolvimento de sistemas inteligentes que imitem aspectos do comportamento humano, como aprendizado, percepção, raciocínio, evolução e adaptação. A descrição do procedimento adotado e das técnicas aplicadas podem ser vistos na Seção Materiais e Métodos.

Este trabalho tem como objetivo analisar uma metodologia para avaliar diferentes técnicas baseadas em inteligência computacional para prever, a partir de um banco de dados pré-existente, a classificação petrográfica de novas lâminas, as quais sejam pertencentes a um mesmo intervalo deposicional. Por meio de medidas de desempenho, são avaliados 4 métodos de classificação em conjunto com outras 5 abordagens de seleção de características, cujo objetivo é melhorar o desempenho dos classificadores.

Materiais e Métodos

Os dados analisados são uma composição de 104 lâminas petrográficas retiradas de 6 poços estratigráficos perfurados na região de Tibagi (Paraná) e Dom Aquino (Mato Grosso do Sul) da Bacia do Paraná. A Bacia do Paraná é uma bacia que se localiza em uma região tectonicamente estável, que possui uma área de 1600000 km^2 com porções localizadas no Brasil, Argentina, Paraguai e no norte do Uruguai.

O procedimento manual realizado por [4] é descrita como segue: as amostras coletadas foram impregnadas para a obtenção de seções delgadas com o objetivo de caracterização do processo diagenético dos arenitos. Esta caracterização gerou dados para interpretar a história diagenética ocorrida e definir a influência da mesma na qualidade dos arenitos como reservatório de hidrocarbonetos. Estas seções delgadas foram analisadas em microscópio Leica DMLP com luz polarizada e refletida e fotografada por câmera digital Nikon Coolpix 9900. Os arenitos foram caracterizados petrograficamente segundo a classificação granulométrica de [16] e textural de [17]. Para a seleção dos grãos foi utilizado o trabalho de [18]. As amostras foram impregnadas com resina epóxi azul para identificação dos poros, segundo o procedimento proposto por [19] e [20]. Para a identificação dos carbonatos foi utilizada a técnica de tingimento. Este procedimento está de acordo com a publicação de [21], a qual orienta para a utilização de uma solução de alizarina (10%) e de ferricianeto em HCl diluído (0.15%). Todos esses procedimentos foram realizados no LGPA da UERJ. As amostras foram analisadas de maneira quantitativa através da contagem de 300 pontos em cada lâmina, com espaçamento de 0.3 mm, onde buscou-se reconhecer as modificações diagenéticas e a relação cronológica entre elas, baseado nas relações texturais e faciológicas observadas. A Figura 1 mostra as etapas do procedimento de identificação e classificação de petrofácies. O método computacional é aplicado após a contagem de 300 pontos em cada lâmina, depois que a base de dados contendo as porcentagens dos constituintes é construída.

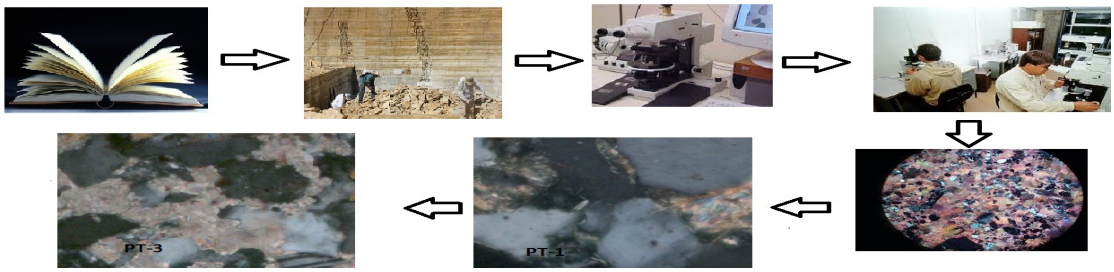

Figura 1: Esquematização do procedimento para realizar a análise de um reservatório através de petrofácies. Este procedimento gerou as amostras usadas neste artigo.

Nas amostras da Bacia do Paraná, foram contabilizadas, para cada lâmina, as porcentagens de 11 constituintes: quartzo, feldspato, glauconita, bioclasto, crescimento secundário de quartzo, caulinita, pirita, siderita, cimento carbonático, porosidade intergranular e porosidade intragranular. De acordo com os resultados obtidos com a análise quantitativa dos dados petrográficos foi possível individualizar as amostras (pela classificação manual) em 9 petrofácies distintas: PT-1, PT-2, PT-3,PT-4, I-1 e I-2 [4] e PG-1, PG-2 e PG-3 [5]. A petrofácies I-1, PG-3 e PT-2 possuem somente uma amostra cada. As petrofácies PT-3, PG-2, PT-4, I-1, PT-1, PG-1 contém 2, 3, 5, 7, 28 e 56 amostras, respectivamente. Percebe-se que as petrofácies PT-1 e PG-1 correspondem a 81% das amostras, resultando em um banco de dados desbalanceados.

A metodologia usada neste trabalho é dividida em dois procedimentos principais: (I) Ajuste dos métodos de classificação através da determinação de seus parâmetros por um processo de busca exaustiva com validação cruzada e (II) uso de métricas para avaliar e comparar os resultados obtidos.

Os classificadores, métodos de validação cruzada e seleção de características utilizados na composição do método proposto foram implementados em linguagem Python, com o auxílio do módulo *Scikit-learn* [13]. Técnicas de classificação tem como objetivo a construção de um modelo (classificador) capaz de predizer a classe de uma nova amostra, com uma resolução satisfatória. A validação cruzada é uma técnica para avaliar a capacidade de generalização de um modelo, através de um conjunto de dados. Procura-se então estimar o quão preciso é este modelo na prática, isto é, o seu desempenho para um novo conjunto de dados. As técnicas de validação cruzada particionam o conjunto de dados em subconjuntos reciprocamente exclusivos, e logo após, usa-se parte destes subconjuntos para a estimação dos parâmetros do modelo (dados de treinamento) e o restante destes são aplicados na validação do modelo (dados de teste). A seleção de características se refere a um procedimento cujo um espaço de dados é convertido em um espaço de características com menor dimensão, mas que ainda armazene a maior parte da informação essencial dos dados. De forma geral, o conjunto de dados sofre uma redução de dimensionalidade.

Aplicou-se métodos de classificação que utilizam distintas abordagens, como K-Nearest Neighbors (KNN), Decision Trees (DT), Linear Support Vector Machine (LinSVM) e Multi-Layer Perceptron (MLP0). O algoritmo KNN [6] é um método não-paramétrico utilizado para a classificação e regressão. Em ambos os casos, a entrada consiste no número de vizinhos mais próximos no espaço de características, ou seja, k. A saída, no caso da classificação, é uma associação de classe. Um objeto é classificado pelo voto da maioria de seus vizinhos, sendo atribuído à classe mais comum entre os seus k vizinhos mais próximos (k é um número inteiro positivo, normalmente pequeno). Se $k = 1$, então o objeto é atribuído à classe do único vizinho mais próximo. DT [7] usam árvores de decisão como um modelo preditivo que mapeia observações sobre um objeto para conclusões sobre a classificação do objeto. É uma das abordagens de modelagem preditiva para utilização na estatística, mineração de dados e aprendizado de máquina. Modelos de árvore em que o rótulo da variável pode tomar um conjunto finito de valores são chamados de árvores de classificação. LinSVM [8] são modelos de aprendizado supervisionado associado com métodos de aprendizado que realizam a análise dos dados e o reconhecimento de padrões. O LinSVM toma como entrada um conjunto de dados e prediz utilizando a abordagem um-contra-todos, para cada entrada, o que o torna mais simples. As redes MLPs [9] têm sido empregadas em diversas áreas, desempenhando tarefas de reconhecimento de padrões, controle e processamento de sinais, entre outras. Uma RNA do tipo MLP é formada por um conjunto de nós fonte, os quais formam a camada de entrada da rede, uma ou mais camadas ocultas e uma camada de saída. Com exceção da camada de entrada, as outras camadas são formadas por neurônios e, dessa maneira, apresentam capacidade computacional. A Figura 2 apresenta um esquema de como os classificadores funcionam.

A estratégia de validação cruzada utilizada foi o Stratified K-Fold (SKF) [10]. A escolha dessa abordagem se deu pelo fato de trabalharmos com dados desbalanceados. No SKF, os subconjuntos são selecionados de modo que o valor médio da resposta seja aproximadamente igual em todos os subconjuntos. Isto significa que cada um dos subconjuntos (treinamento e teste) contém aproximadamente as mesmas proporções dos tipos de classe.

Para a seleção de características fez-se uso de técnicas baseadas no teste de estatística univariada. Os seguintes métodos foram empregados: Select K-Best (SKB), Select Percentile (SP), Select Fpr (SFpr) e Generic Univariate Select (GUS) [11]. O SKB seleciona as k características com maior pontuação. No SP as características são escolhidas de acordo com o percentil das maiores pontuações. Já o SFpr escolhe através dos p-valores abaixo de α (α geralmente 0.05) baseado no teste

FPR. O Teste FPR (False Positive Rate) controla a quantidade total de falsas detecções. O GUS é um método univariado com estratégia configurada. Isto permite selecionar a melhor estratégia de seleção univariada com hiper-parâmetro de busca estimador.

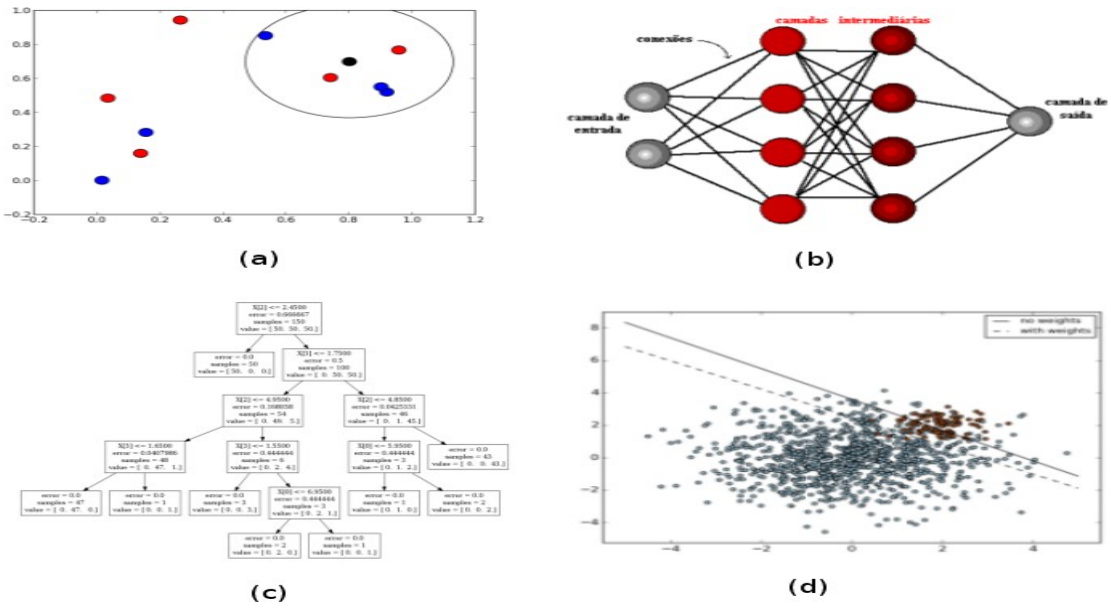

Figura 2: Esquema dos classificadores usados neste artigo: (a) Métodos dos Vizinhos mais Próximos (KNN) (b) Rede neuronal do tipo multilayer perceptron (MLP0) – 2 variáveis de entrada, 2 camadas ocultas e 1 camada de saída (c) Árvores de decisão (DT) e (d) Máquinas de Vetores Suporte com funções lineares (LinSVM)

Foi criada uma assembleia de constituintes, uma divisão dos constituintes em Detríticos e Diagenéticos. Detríticos são os grãos depositados, provenientes da desagregação da litologias da área fonte e/ou de altos internos da bacia. Diagenéticos são os constituintes precipitados quimicamente depois da deposição dos sedimentos, durante o processo diagenético. Assim, esses constituintes visam representar nas amostras os processos de formação do arcabouço e de consolidação. Os minerais considerados Detríticos foram: quartzo, feldspato e bioclasto. Já como Diagenéticos foram: glauconita, crescimento secundário de quartzo, caulinita, pirita, siderita, cimento carbonático, porosidade intergranular e porosidade intragranular. Efetuou-se a normalização desses dados separadamente, reduziu a dimensionalidade para duas, utilizando Análise de Componentes Principais (PCA) e depois agrupou-se os mesmos, formando uma base de dados. O intuito dessa abordagem é verificar se essa separação dos constituintes influencia na classificação dos dados.

Realizou-se a avaliação dos métodos de classificação juntamente as abordagens citadas a partir das métricas acurácia e F1 [12]. A acurácia mede a porcentagem de acertos do método, comparando as petrofácies encontradas com as classificadas pelo método manual, através da contagem direta. O F1 é dado por 2tp/(2tp + fp + fn), onde tp é taxa de verdadeiro positivo, fn é a taxa de falso negativo e fn é a taxa de falso negativo.

Resultados e Discussão

A Figura 3 apresenta o gráfico de barras com a acurácia e o F1 encontrados através da emprego dos métodos de seleção de características (SKB, SP, SFpr e GUS), da assembleia de constituintes (ASBLY) e sem o uso de nenhuma abordagem (NFS) além dos métodos de classificação combinados com o SKF.

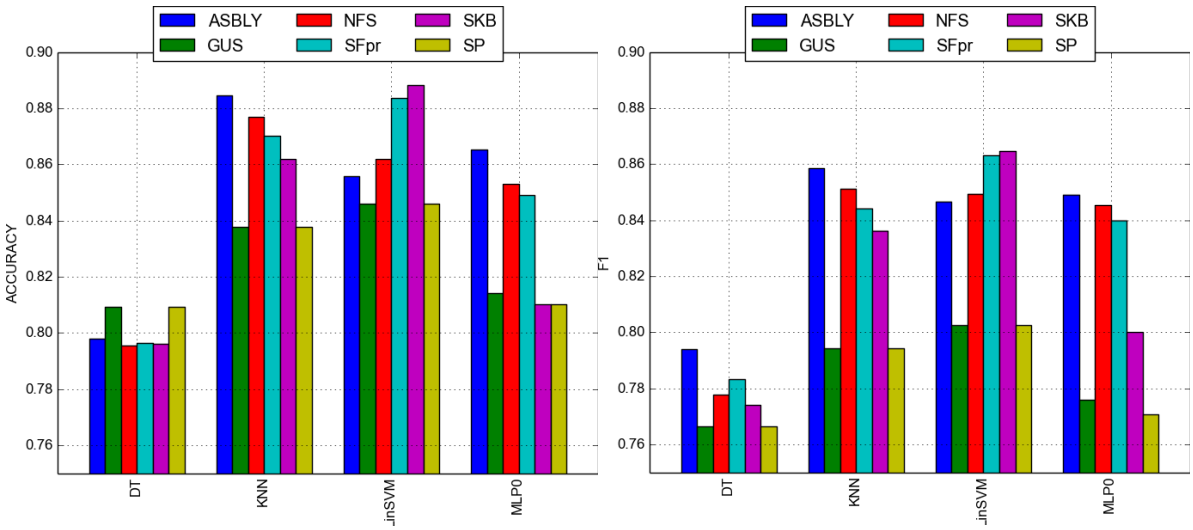

Figura 3: Acurácia e F1– 30 execuções. Foi utilizado k=5 para o SKF, k-1 conjuntos para treinamento e o restante para teste. Para os classificadores MLP0 e KNN o uso do ASBLY resultou em uma acurácia mais próxima de 1.0. Nenhum dos outros seletores de características obtiveram bom desempenho para estes dois classificadores, o segundo melhor resultado foi para o NFS. Para o LinSVM, o método SKB gerou uma acurácia mais alta. No DT, o melhor resultado foi com o GUS e o SP, que obtiveram o mesmo valor da acurácia. Para os métodos de classificação MLP0, KNN e DT o uso da assembleia de constituintes resultou em um F1 mais alto. Já para o LinSVM, o SKB originou melhor F1.

O bom desempenho do método MLP0 utilizando a assembleia de constituintes pode ser explicada pelo fato da complexidade da rede ter sido reduzida, uma vez que foram utilizadas quatro características, sendo duas detríticas e duas diagenéticas. O MLP0 é uma rede neuronal de múltiplas camadas, e a redução de dimensionalidade proporciona a redução da quantidade de conexões na arquitetura e na rede e consequentemente facilita a convergência para a solução no processo de minimização do funcional do erro médio quadrático. Neste caso, a perda de informações devido a redução de dimensões do problema é compensada com a simplificação do problema de otimização envolvido no treinamento da rede neuronal. O seletor nesse caso melhorou o desempenho do classificador em 1.2% em relação ao NFS.

No caso do LinSVM o SKB apresentou melhor acurácia e F1. Foram selecionadas as 5 características com maior pontuação. O seletor nesse caso melhorou o desempenho do classificador em 0.1% em relação ao Sfpr e em 2% em relação ao NFS, essas abordagens tiveram desempenho semelhantes. Nota-se que a assembleia de constituintes não se destacou, pois nesse classificador os dados sofrem uma transformação não linear, dessa maneira, houve duas transformações, uma realizada pela assembleia de constituintes e em seguida outra realizada pela formulação matemática do classificador LinSVM, o que ocasionou uma perda de informação e resultou no baixo desempenho do método.

Para o KNN com 4 vizinhos utilizando a assembleia de constituintes obteve melhor resultado. A abordagem nesse caso melhorou o desempenho do classificador em 0.8% em relação ao NFS, segunda abordagem com melhor desempenho. Um dos fatores que influenciou os bons resultados da abordagem utilizada foi a normalização dos dados que não ocorre nos outros métodos (SKB, SP, SFpr, GUS e NFS). Por outro lado, O DT não obteve bons resultados comparado com os outros métodos independente da técnica de seleção de características empregada. As árvores de decisão criam árvores tendenciosas se algumas classes possuem mais amostras que outras. A base de dados em questão está desbalanceada, e o uso do SKF não foi suficiente para que o classificador gerasse bons resultados.

Conclusões

Com base nos experimentos realizados nesse trabalho, o método LinSVM combinado com o SKB obteve melhor resultado. O método de classificação que resultou em valores mais baixos de acurácia e F1 foi o DT. Entre as abordagens empregadas, o método GUS e o SP não mostraram bom desempenho na resolução do problema proposto, independente do classificador em que foram empregados. O uso de assembleia de constituintes teve um bom desempenho, mostrando ser uma boa estratégia. A ferramenta computacional desenvolvida permite obter resultados semelhantes àqueles obtidos pelo método convencional de individualização, a acurácia dos métodos variou entre 77% a 89% atingindo uma boa precisão na tarefa proposta. O resultado foi comprometido devido ao desbalanceamento dos dados. Como trabalhos futuros pretende-se aplicar técnicas de balanceamento de dados e analisar o impacto destes na metodologia proposta.

Agradecimentos

Os autores agradecem à Universidade Federal de Juiz de Fora (UFJF), ao Programa de Pós-Graduação em Modelagem Computacional (PPGMC), à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes) pelo suporte dado e à Universidade do Estado do Rio de Janeiro (UERJ) por conceder acesso às lâminas utilizadas para análise e obtenção dos dados.

Referências Bibliográficas

[1] Cevolani et al. Computational Methodology to Study Heterogeneities in Petroleum Reservoirs. EAGE Annual Conference & Exhibition / SPE EUROPEC, 2013.

[2] Cunha, E. S. 2002. Identificação de litofácies de poços de petróleo utilizando um método Baseado em Redes Neurais Artificiais. Dissertação de Mestrado. Programa de Pós-Graduação em Informática.UFCG.

[3] Lim, J. S. Reservoir properties determination using fuzzy logic and neural networks from well data in offshore Korea. Journal of Petroleum Science and Engineering, v.49, p.182-192, 2005.

[4] Oliveira, L. C. 2009. Estudo das relações entre o arcabouço estratigráfico e as alterações diagenéticas observadas na seção Devoniana da bacia do Paraná. Dissertação de Mestrado. Faculdade de Geologia – UERJ. Brasil.

[5] Brazil, F. A. F. 2004. Estratigrafia de sequências e processo diagenético: exemplos dos arenitos marinho-rasos da Formação Ponta Grossa, Noroeste da Bacia do Paraná. Dissertação de Mestrado. Faculdade de Geologia – UERJ. Brasil.

[6] Andoni A. & Indyk P. Near-Optimal Hashing Algorithms for Approximate Nearest Neighbor in High Dimensions. Foundations of Computer Science, 2006. FOCS '06. 47th Annual IEEE Symposium.

[7] M. Dumont *et al*. Fast multi-class image annotacion with random subwindows and multiple output randomized trees, International Conference on Computer Vision Theory and Applications 2009.

[8] Lorena A. C. & de Carvaho, A.C. P. L. F. 2003. Introdução às Máquinas de Vetores Suporte. Relatórios Técnicos do ICMC. Instituto de Ciências Matemáticas e de Computação.

[9] Affonso E.T. F. *et al*. Uso de Redes Neurais Multilayer Perceptron (MLP) em um sistema de bloqueio de websites baseado em conteúdo. Mecánica Computacional.v. XXIX, p. 9075-9090, 2010.

[10] Kohavi, R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. IJCAI'95 Proceedings of the 14th International Joint Conference on Artificial Intelligence, v.2, p.1137-1143, 1995.

[11] Pedregosa *et al.* 1.13. Feature selection. Disponível em:< http://scikit-learn.org/stable/modules/feature_selection.html>. Acesso em: 02/06/2015.

[12] Powers, D. M. W. Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. Technical Report SIE-07-001, December, 2007.

[13] Pedregosa *et al.* Scikit-learn: Machine Learning in Python, JMLR, v. 12, p. 2825-2830, 2011.

[14] Nurcihan Ceryan, Application of support vector machines and relevance vector machines in predicting uniaxial compressive strength of volcanic rocks, Journal of African Earth Sciences, Volume 100, December 2014, p 634-644.

[15] da Silva, P. N., Gonçalves , E. C., Rios, E. H., Muhammad, A., Moss, A., Pritchard, T., Glassborow, B., Plastino, A., Automatic classification of carbonate rocks permeability from 1H NMR relaxation data, Expert Systems with Applications, Volume 42, Issue 9, 1 June 2015, p. 4299-4309.

[16] Wentworth, C.K. 1922. A Scale of grade and class terms for clastic sediments, J. Geol., 30, 377–392.

[17] Folk, R.L. 1980. Petrology of sedimentary rocks. Hemphill Publishing Company. Austin, Texas, EUA.184p.

[18] Beard, D. C. & Weil, P. K., 1976. Influence on porosity of unconsolidated sand. AAPG Bulletin, 57(2): 349-369.

[19] De Césero, P., Mauro, L.M., De Ross, L.F. 1989. Técnicas de preparação de lâminas petrográficas e de moldes de poros na Petrobrás. Boletim de Geociências, Petrobrás, 3(1/2): 105-106.

[20] Lindholm, R. C., 1972. Calcite staining: semiquantitative determination of ferrous iron. Jornal of Sedimentary petrology, 42: 239-242.

[21] Evamy B. D., 1963. The application of chemical staining technique to a study of dedolomitization. Sedimentology, 2: 164-170.