DISCRIMINAÇÃO LITOLÓGICA POR ATRIBUTOS SÍSMICOS ELÁSTICOS: UMA ABORDAGEM POR SISTEMAS FUZZY-GENÉTICOS

Eric da Silva Praxedes
PUC-Rio - Rua Marquês de São Vicente, 225, Gávea - Rio de Janeiro/RJ
praxedes@ele.puc-rio.br

Adriano Soares Koshiyama
PUC-Rio - Rua Marquês de São Vicente, 225, Gávea - Rio de Janeiro/RJ
adriano@ele.puc-rio.br

Marley Maria Bernardes Rebuzzi Vellasco
PUC-Rio - Rua Marquês de São Vicente, 225, Gávea - Rio de Janeiro/RJ
marley@ele.puc-rio.br

Marco Aurélio Cavalcanti Pacheco
PUC-Rio - Rua Marquês de São Vicente, 225, Gávea - Rio de Janeiro/RJ
marco@ele.puc-rio.br

Ricardo Tanscheit
PUC-Rio - Rua Marquês de São Vicente, 225, Gávea - Rio de Janeiro/RJ
ricardo@ele.puc-rio.br

Resumo: Este trabalho propõe uma metodologia para discriminação litológica de novas jazidas de petróleo a partir do uso de Sistemas Fuzzy-Genéticos, em destaque o modelo GP-FIS (Genetic Programming for Fuzzy Inference System). A grande vantagem da modelagem proposta é possibilitar identificar, a partir de padrões sísmicos, o tipo de rocha de uma determinada região sem a necessidade de abrir novos poços. Assim, busca-se um modelo com boa acurácia, aprendizado automático e que proporcione duas flexibilidades aos especialistas: avaliar o grau de pertinência de um determinado padrão sísmico aos diferentes tipos de rocha e avaliação a nível linguístico da resposta do modelo. Assim, a ferramenta final elaborada proporciona apoio à decisão como também extração e descoberta de conhecimento. Além do modelo GPFIS são avaliados 7 outras metodologias para classificação, através de dados de um poço da costa brasileira.

Palavras-Chave: Classificação; Litologia; Óleo & Gás; Petróleo; Sistemas Fuzzy-Genéticos.

Abstract: This work proposes a new methodology for lithological discrimination, using GPFIS model (Genetic Programming for Fuzzy Inference System) a Genetic Fuzzy System based on Multi-Gene Genetic Programming. The main advantage of our approach is the possibility to identify, through seismic patterns, the rock types in new regions without requiring opening wells. Thus, we seek for a reliable model that provides two flexibilities for the experts: evaluate the membership degree of a seismic pattern to the several rock types and the chance to analyze at linguistic level the model output. Therefore, the final tool must afford knowledge discovery and support to the decision maker. Also, we evaluate other 7 classification models (from statistics and computational intelligence), using a database from a well located in Brazilian coast.

Keywords: Classification, Lithology, Oil & Gas, Genetic Fuzzy Systems.

1. INTRODUÇÃO

Uma das tarefas mais importantes na indústria de exploração e produção de petróleo é a identificação litológica. Litologia é a descrição das características físicas macroscópicas de uma rocha tais como cor, textura, tamanho do grão e conteúdo mineral [1,2]. Com base nessa descrição, e conhecendo-se a localização de cada tipo de rocha no poço, é possível inferir onde se encontram as formações geradoras de contenção do hidrocarboneto e principalmente o reservatório, necessários para a ocorrência de um sistema petrolífero.

Para esta tarefa existem diversas fontes de informação, mas uma das principais para subsidiar a identificação litológica é a perfilagem. Esta consiste em medidas físicas colhidas por ferramentas que são baixadas dentro do poço. Devido às ferramentas de perfilagem medirem propriedades das rochas no subsolo, seus registros são intrinsecamente geológicos [3]. Tradicionalmente, técnicas estatísticas são utilizadas para proporcionar a identificação litológica através do estudo dos perfis. Uma das mais utilizadas é a Análise Discriminante [4]. Mais recentemente, técnicas de Inteligência Computacional vêm sendo utilizadas com relativo sucesso. Destacam-se o uso de Redes Neurais, Máquina de Vetores de Suporte e Sistemas de Inferência Fuzzy [5,6,7,8].

Contudo, na maioria dos trabalhos os perfis utilizados na identificação litológica são aqueles disponíveis nos poços, tais como os de raios gama, porosidade neutrônica e resistividade. Estas informações são somente conhecidas após a abertura dos poços. Na fase anterior, ou seja, fora do poço, os únicos registros disponíveis da subsuperfície são os atributos derivados dos levantamentos sísmicos.

Para que modelos de identificação de litologia possam ser aplicados fora dos poços, faz-se necessária a utilização de características que estejam presentes tanto nos poços como fora deles. Nos poços, onde a litologia é conhecida através dos perfis convencionais e de outras informações, os modelos são treinados e, em seguida, testados para aferir a sua acurácia. Fora dos poços, utilizando-se as mesmas características, os modelos são então aplicados para que a qualidade de seu resultado seja avaliada por um geólogo ou geofísico.

Para tanto, tal modelo deve ir além tanto em termos de acurácia da abordagem padrão (Análise Discriminante Linear), assim como oferecer aos especialistas interpretações dos resultados obtidos. Uma alternativa viável é partir do uso de Sistemas Fuzzy-Genéticos [9,10], que conferem ao seu usuário relativa acurácia e boa compreensão linguística das classificações efetuadas pelo modelo. A partir do modelo GPFIS (Genetic Programming Fuzzy Inference System) [11], este trabalho tem por objetivo avaliar a qualidade desta abordagem frente a técnicas de classificação da Inteligência Computacional e Estatística.

Este trabalho está assim organizado: a próxima seção descreve noções sobre Física de Rochas e apresenta de forma aprofundada o problema enfrentado. A seção 3 dispõe o modelo GPFIS adequado para o problema de classificação enunciado na seção 2. A seção 4 disserta sobre os demais modelos de classificação usados, as métricas de avaliação, o procedimento experimental, assim como dos resultados e discussões. Por fim, a seção 5 apresenta as considerações finais e os trabalhos futuros.

2. NOÇÕES SOBRE FÍSICA DE ROCHAS E PROBLEMA ABORDADO

A teoria de física de rochas é a área da geofísica que fornece as relações entre os atributos elásticos sísmicos, medidos a partir da superfície da Terra, de dentro dos poços ou em laboratório, com as propriedades petrofísicas das rochas [12]. Ela fornece o entendimento e as ferramentas teóricas para aperfeiçoar a caracterização baseada em dados elásticos.

Dentre os atributos elásticos sísmicos existentes, um dos mais utilizados é a impedância acústica, tanto a compressional (IP) como a cisalhante (IS). Ela é definida como o produto entre a densidade e a velocidade de propagação da onda compressional e cisalhante em um

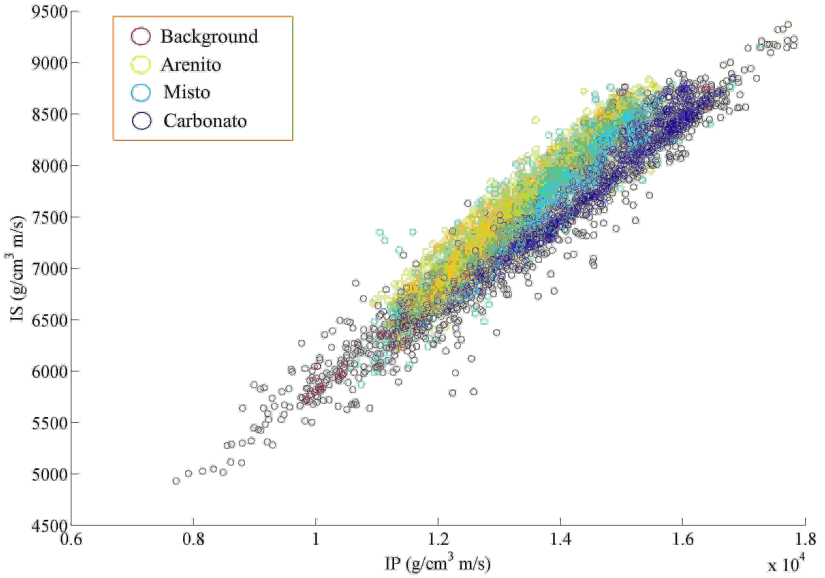

Figura 1: Relacionamento entre IP e IS e os tipos de rochas.

meio [13,14]. As impedâncias podem ser calculadas tanto dentro do poço como fora dele. No poço, elas podem ser calculadas através do produto entre o perfil de densidade e o perfil que mede o tempo de trânsito compressional e cisalhante. Fora do poço, é possível produzir um volume com valores de impedância através do processo denominado inversão sísmica, que utiliza a equação de Aki-Richards e os angle-gathers para obter tais volumes [12]. Utilizando-se as impedâncias compressional e cisalhante, é possível criar um modelo com dados de poços para classificar litologias e extrapolá-lo para os dados de sísmica.

Para este trabalho foram utilizados dados de um poço da costa brasileira. O poço selecionado para este estudo localiza-se em uma região com um sistema misto para a formação de rochas sedimentares. A litologia encontrada neste poço foi interpretada por um geólogo especialista que, para tal, utilizou-se dos perfis convencionais de poços de raios gama, densidade, fator fotoelétrico e sônico compressional, além da descrição petrográfica de amostras laterais e de fácies dos perfis de imagem (resistivo e acústico). Nesta interpretação litológica foram encontradas sete diferentes tipos de rochas: arenito, arenito carbonático, calcarenito, mudstone, packstone, argilito arenoso e folhelho. Estes foram agrupados em quatro diferentes classes, pois as características que diferenciam as rochas de uma mesma classe não são detectadas na escala sísmica. Na Tabela 1 é apresentado o agrupamento realizado para criar as quatro classes.

Tabela 1: Tipos de litologias e agrupamentos usados no poço em análise.

Litologia	Classe	Padrões	Frequência
Arenito	Arenito	1991	33,21%
Arenito Carbonático	Misto	1803	30,08%
Calcarenito Mudstone Packstone	Carbonato	1480	24,69%
Argilito Arenoso Folhelho	Background	721	12,03%

Para realizar a classificação serão usados, além dos valores diretos das impedâncias (IP e IS), mais dois atributos calculados a partir delas. O primeiro é a diferença entre IP e IS (IP – IS) enquanto o segundo é a Razão de Poisson (RP): $RP = \frac{(IP^2 - 2IS^2)}{2(IP^2 - IS^2)}$. A escolha desses

outros atributos foi baseada no estudo de física de rochas, que aponta ambos como bons discriminantes litológicos [13,15]. A Figura 1 apresenta o gráfico entre IP e IS.

A partir dos atributos IP, IS, IP-IS e RP amostrados busca-se um modelo que possa inferir, a partir destes comportamentos, o tipo de rocha em questão: Arenito, Misto, Carbonato ou Background. A grande vantagem do GPFIS é a possibilidade de oferecer tal interpretação em nível linguístico e ainda possibilitar que o especialista decida qual o tipo de rocha, a partir do grau de pertinência de uma amostra sísmica às diferentes classes.

3. MODELO GPFIS

Esta seção exibe o modelo GPFIS [11], dedicadamente formulado para solucionar problemas de classificação. Como o modelo se baseia na Programação Genética Multigênica [16, 17], a primeira parte da seção exibe esta variante da Programação Genética clássica, sendo após formulado o modelo GPFIS.

3.1. Programação Genética Multigênica

A Programação Genética (PG) [18] é uma técnica da Computação Evolutiva inspirada nos conceitos de seleção natural e recombinação genética. A Programação Genética Multigênica (PGMG) [16,17] pode ser encarada como uma generalização da PG tradicional, pois denota um indivíduo como um complexo de estrutura em árvores (funções), que, da mesma forma que na PG, recebe um conjunto de Terminais X_j (atributos em reconhecimento de padrões, defasagens de séries temporais, etc.), buscando predizer a saída Y. A representação da PGMG é similar ao da PG no tocante à estrutura em árvore, porém um indivíduo para a PGMG é um complexo de estruturas em árvore (Figura 2).

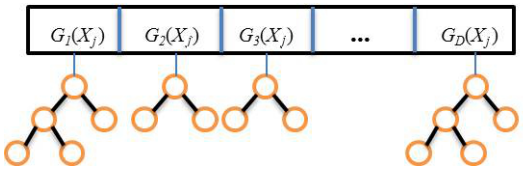

Figura 2: Exemplo de um indivíduo multigênico.

Cada árvore desta estrutura é uma solução parcial para o problema. É fácil ver que quando $D = 1$, a PGMG reduz-se à solução obtida por uma PG clássica ($d = 1, ..., D$). O processo de avaliação e seleção são efetuados de forma similar a PG. Com relação aos operadores de recombinação, a operação de mutação na PGMG é similar à efetuada na PG clássica. No caso da operação de cruzamento, é necessária uma distinção no nível em que a operação é realizada, sendo possível aplicar o cruzamento no baixo e no alto níveis. O baixo nível é o espaço onde é possível manipular as estruturas (Terminais e Operações Matemáticas) das equações presentes em um indivíduo. No caso, tanto a mutação quanto o cruzamento de baixo nível na PGMG são semelhantes ao modo efetuado na PG.

Um exemplo de cruzamento de alto nível é apresentado na Figura 3. O alto nível é o espaço que se manipula de forma macro as equações presente no indivíduo. Logo, verifica-se que, a partir de dois pontos aleatórios, são permutadas equações de um indivíduo para o outro. Os efeitos do cruzamento de alto nível tendem a afetar mais substancialmente a saída resultante do que a operação de cruzamento de baixo nível e a mutação.

Em linhas gerais, o procedimento evolutivo da PGMG se diferencia da PG pela adição de dois parâmetros: Número máximo de árvores por indivíduo e Taxa de Cruzamento de Alto Nível. No caso do número máximo de árvores por indivíduo, sempre se utiliza um valor elevado para que não haja empecilhos no processo de sintetização da solução. Com respeito

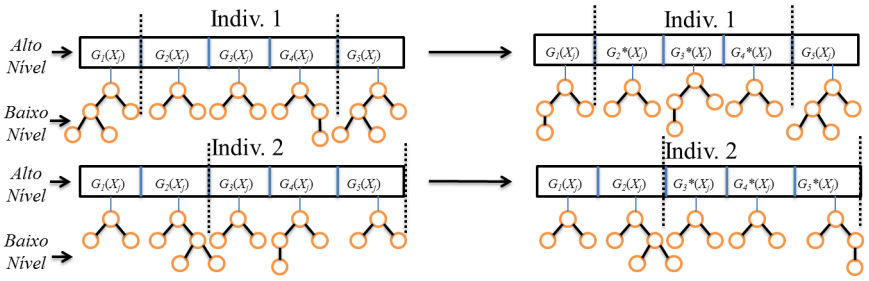

Figura 3: Operação de cruzamento de alto nível.

à taxa de cruzamento de alto nível, esta é um parâmetro que deve ser previamente definido, sendo sempre seu valor apresentado na tabela de configurações do algoritmo.

3.2. Modelo GPFIS

Esta seção aborda o modelo Genetic Programming Fuzzy Inference System (GPFIS) para classificação [11]. São descritas suas etapas de construção, desde o mapeamento de valores precisos em graus de pertinência a conjuntos fuzzy,o procedimento de Inferência que é subdivido em Formulação, Particionamento e Agregação. Após o processo de Inferência, é realizada a Decisão e Avaliação.

3.2.1. Fuzzificação

Em classificação, a principal informação disponível são os n padrões $\mathbf{x_p} = [x_{p1}, x_{p2}, ..., x_{pK}]$ dos K atributos X_k presentes na base de dados ($p = 1, ..., n$ e $k = 1, ..., K$). A informação dos n padrões são usadas para distinguir a qual h-ésima classe um novo padrão $\mathbf{x_p}^*$ pertence ($h = 1, ..., H$). A etapa de fuzzificação estabelece os conjuntos fuzzy A_{jk} associados a cada k-ésimo atributo. No caso em estudo os atributos são: $X_1 = \{IP\}$, $X_2 = \{IS\}$, $X_3 = \{IP - IS\}$ e $X_4 = \{RP\}$.

A etapa de fuzzificação leva em conta três fatores: forma funcional, definição do suporte de cada função de pertinência $\mu_{A_{jk}}(x_{pk})$ e rótulo linguístico apropriado, qualificando o subespaço compreendido pela função de pertinência com um adjetivo correspondente ao contexto. Após uma discussão com o corpo de especialistas no tema, a disposição das funções de pertinência para cada X_k são dadas pela Figura 4.

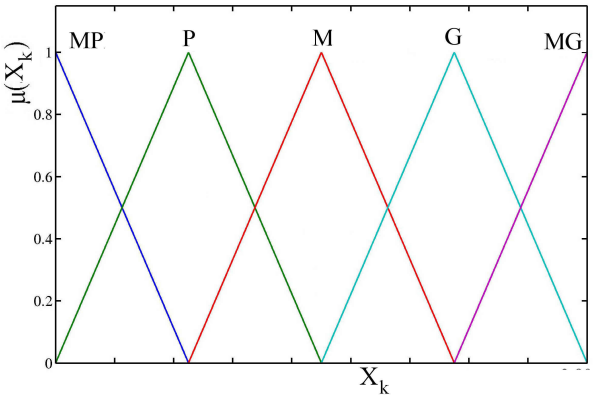

Figura 4: Funções de pertinência para as variáveis X_k.

A partir da fuzzificação de cada padrão, a etapa de Inferência utiliza a informação contida em cada A_{jk} para melhor predizer a classe de $\mathbf{x_p}^*$.

3.2.2. Inferência

3.2.2.1. Formulação

De modo sucinto, o modelo GPFIS busca um conjunto de funções que satisfaçam a seguinte representação:

$$\mu_{C_p \in 1}(\mathbf{x_p}) = g[f_{d \in s_1}(\mu_{A_{j1}}(x_{p1}), ..., \mu_{A_{jK}}(x_{pK})), ..., f_{d \in s_1}(\mu_{A_{j1}}(x_{p1})..., \mu_{A_{jK}}(x_{pK}))] + \varepsilon_{p1} \Rightarrow$$
$$\mu_{C_p \in 1}(\mathbf{x_p}) = \hat{\mu}_{C_p \in 1}(\mathbf{x_p}) + \varepsilon_{p1} \quad (1)$$
$$...$$
$$\mu_{C_p \in H}(\mathbf{x_p}) = g[f_{d \in s_H}(\mu_{A_{j1}}(x_{p1}), ..., \mu_{A_{jK}}(x_{pK})), ..., f_{d \in s_H}(\mu_{A_{j1}}(x_{p1}), ..., \mu_{A_{jK}}(x_{pK}))] + \varepsilon_{pH} \Rightarrow$$
$$\mu_{C_p \in H}(\mathbf{x_p}) = \hat{\mu}_{C_p \in H}(\mathbf{x_p}) + \varepsilon_{pH} \quad (2)$$

Os elementos pertencentes a cada expressão são descritos a seguir:

- $f_{d \in s_h}(\mu_{A_{j1}}(x_{p1})..., \mu_{A_{jK}}(x_{pK}))$: a função $f_{d \in s_h} : [0,1]^K \rightarrow [0,1]$, $d = 1, ..., D$ descreve a forma de relacionamento das funções de pertinência de cada k-ésimo atributo com respeito a h-ésima classe. Cada $f_{d \in s_h}(\mu_{A_{j1}}(x_{p1})..., \mu_{A_{jK}}(x_{pK}))$ descrevem uma regra fuzzy, a partir de operadores t-norma, t-conorma, negação, etc., com o propósito de representar conectivos lógicos ("e", "ou"e "não") e modificadores linguísticos ("muito"e "pouco", por exemplo). O índice $d \in s_h$ é melhor descrito na seção Particionamento, mas em linhas gerais indica quais d-ésimas funções (premissa de regra) estão relacionadas a h-ésima classe.

- $g[f_{d \in s_h}, ..., f_{d \in s_h}]$: A função $g : [0,1]^{card(s_h)} \rightarrow [0,1]$ é um operador de agregação, cujo papel é unir os graus de ativação relativos ao conjunto de regras associadas a cada h-ésima classe em um valor final.

- $\mu_{C_p \in h}(\mathbf{x_p})$ mensura o grau de pertinência de $\mathbf{x_p}$ à h-ésima classe. É sempre $\{0, 1\}$, ou seja, o p-ésimo padrão pertence ou não a classe h.

- $\hat{\mu}_{C_p \in h}(\mathbf{x_p})$ mede o grau de pertinência estimado de $\mathbf{x_p}$ à h-ésima classe e assume valores entre $[0, 1]$.

- ε_{ph}: desvio entre o observado $\mu_{C_p \in h}(\mathbf{x_p})$ e o estimado $\hat{\mu}_{C_p \in h}(\mathbf{x_p})$.

O objetivo do modelo GPFIS é buscar as $f_{d \in s_h}(\mu_{A_{j1}}(x_{p1})..., \mu_{A_{jK}}(x_{pK}))$ de modo que produzam uma estimativa $\hat{\mu}_{C_p \in h}(\mathbf{x_p})$ que minimize $\sum_{p=1}^{n} |\varepsilon_{ph}|$. Para tanto, o modelo GPFIS utiliza elementos da PGMG, de modo a sintetizar o conjunto de premissas de regras $f_d(\mu_{A_{j1}}(x_{p1})..., \mu_{A_{jK}}(x_{pK}))$. A partir da disponibilidade de um conjunto de premissas $f_d(\mu_{A_{j1}}(x_{p1})..., \mu_{A_{jK}}(x_{pK}))$, torna-se necessário definir uma classe consequente (isto é, tornar $f_d(\mu_{A_{jk}})$ em $f_{d \in s_h}(\mu_{A_{j1}}(x_{p1})..., \mu_{A_{jK}}(x_{pK}))$). As técnicas de Particionamento, assunto do próximo tópico, são os mecanismos que podem ser usados para a escolha da classe mais bem associada a cada $f_d(\mu_{A_{j1}}(x_{p1})..., \mu_{A_{jK}}(x_{pK}))$.

3.2.2.2. Particionamento

Seja $d = 1, ..., D$ o conjunto de índices das funções $f_d(\mu_{A_{jk}})$ e $S = \{s_0, s_1, s_2, ..., s_H\}$ o conjunto das partes de d, onde s_h são os índices das f_d destinadas a h-ésima classe. Por fim, s_0 é o conjunto das $f_d(\mu_{A_{j1}}(x_{p1})..., \mu_{A_{jK}}(x_{pK}))$ direcionadas para nenhuma classe específica (isto é, os antecedentes descartados). O método da Máxima Confiança avalia o grau de compatibilidade da parte antecedente $f_d(\mu_{A_{jk}})$, com respeito a todas as H classes. Isto é,

deseja-se definir o consequente de regra mais confiável para uma dada premissa. Assim, para cada uma das H classes é computado um Grau de Confiança à Classe h (CD_h), dado por:

$$CD_h = \frac{\sum_{p \in h} f_d(\mu_{A_{j1}}(x_{p1})..., \mu_{A_{jK}}(x_{pK}))}{\sum_{p=1}^{n} f_d(\mu_{A_{j1}}(x_{p1})..., \mu_{A_{jK}}(x_{pK}))} \tag{3}$$

o CD_h pode ser avaliado como a identificação da parte antecedente aos padrões da classe h, com relação ao total de compatibilidade da parte antecedente à classe h e às demais. Logo, $0 \leq CD_h \leq 1$, onde $CD_h = 1$ significa compatibilidade total, e $CD_h = 0$, o contrário. A definição da classe C das $f_d(\mu_{A_{jk}})$ é dada pelos passos, para toda $d = 1, ..., D$ premissa de regra:

1. Calcula-se a CD_h para todos as H classes.
2. A h-ésima classe para a $f_d(\mu_{A_{j1}}(x_{p1})..., \mu_{A_{jK}}(x_{pK}))$ é definida pela que maximizar CD_h.
3. Insere-se o índice da $f_d(\mu_{A_{j1}}(x_{p1})..., \mu_{A_{jK}}(x_{pK}))$ no respectivo s_h.
4. Caso a $f_d(\mu_{A_{j1}}(x_{p1})..., \mu_{A_{jK}}(x_{pK}))$ tenha $CD_h = 0$ para todo h, insere-se o índice desta em s_0.

Logo, nem toda $f_d(\mu_{A_{j1}}(x_{p1})..., \mu_{A_{jK}}(x_{pK}))$ estará associada a um determinado consequente, como também pode haver algum consequente que esteja inativo. Após a definição dos $f_{d \in s_h}(\mu_{A_{j1}}(x_{p1})..., \mu_{A_{jK}}(x_{pK}))$, isto é, a base de regras fuzzy, é possível avaliar para qual classe um $\mathbf{x_p}$ é mais pertinente. Como uma mesma classe pode possuir diversas regras associadas, torna-se necessário agregar as ativações provenientes da compatibilidade de $\mathbf{x_p}$ para cada regra, de modo a gerar uma estimativa final $\hat{\mu}_{C_p \in h}(\mathbf{x_p})$.

3.2.2.3. Agregação

Na área de Sistemas Fuzzy-Genéticos é comum usar o operador de agregação Máximo. Contudo, na aplicação em estudo agrega-se as regras referentes ao mesmo consequente pelo operador de combinação convexa, isto é:

$$g[f_{d \in s_h}(\mu_{A_{j1}}(x_{p1})..., \mu_{A_{jK}}(x_{pK}))] \rightarrow \sum_{d \in s_h} w_{d \in s_h} f_{d \in s_h}(\mu_{A_{jk}}), \quad \text{com} : \sum_{d \in s_h} w_{d \in s_h} = 1, w_{d \in s_h} \geq 0 \tag{4}$$

este operador generaliza a média aritmética na medida que os pesos $w_{d \in s_h}$ podem ser quaisquer valores entre [0,1], com a restrição que somem 1. Semelhantemente, a interpretação se altera, tal que $w_{d \in s_h}$ indica o grau de influência dessa regra no resultado final. Após a etapa de agregação, são obtidos cada $\hat{\mu}_{C_p \in 1}(\mathbf{x_p}), ..., \hat{\mu}_{C_p \in H}(\mathbf{x_p})$. Contudo, é necessário definir de maneira crisp a classe associada ao p-ésimo padrão.

3.2.3. Decisão

Para a formulação explorada, a Decisão pelo pertencimento do p-ésimo padrão $\mathbf{x_p}$ à classe $h = 1, ..., H$ é dada por:

$$\hat{C}_p(\mathbf{x_p}) = \arg_h \max\{\hat{\mu}_{C_p \in 1}(\mathbf{x_p}), ..., \hat{\mu}_{C_p \in H}(\mathbf{x_p})\} \tag{5}$$

onde $\hat{C}_p(\mathbf{x_p})$ é a classe estimada, resultado do h-ésimo argumento que assume o valor máximo na expressão (5). A ideia por trás é indicar que $\mathbf{x_p}$ pertence à classe com a qual é mais compatível, segundo as regras disponíveis. Quando há empate, uma heurística decisória pode ser aplicada (a classe que possui maior proporção), ou nenhuma classe específica é atribuída a $\mathbf{x_p}$.

3.2.4. Avaliação

De forma sucinta, a avaliação no modelo GPFIS é definida por um objetivo primário, minimização do erro, e secundário, redução da complexidade do indivíduo. O objetivo primário domina a forma de posicionamento dos indivíduos da população, enquanto que o segundo se manifesta como critério de desempate.

A função de avaliação para problemas de classificação é dada pelo Erro Médio de Classificação (EMC) como:

$$EMC = \frac{\sum_{p=1}^{n} |C_{p\in h}(\mathbf{x_p}) - \hat{C}_{p\in h}(\mathbf{x_p})|}{n} \tag{6}$$

onde para dado um padrão $\mathbf{x_p}$, $|C_{p\in h}(\mathbf{x_p}) - \hat{C}_{p\in h}(\mathbf{x_p})| = 0$, se $C_{p\in h}(\mathbf{x_p}) = \hat{C}_{p\in h}(\mathbf{x_p})$ e 1, caso contrário. O indivíduo que minimizar o EMC é considerado o melhor na população.

O segundo objetivo é a redução da complexidade. Esta é baseada no método de Pressão Lexicográfica Parcimoniosa [19]. A ideia por de trás do método é: dado dois indivíduos com desempenhos idênticos, o melhor entre eles é o que possui menor número de nós na árvore. Isto indica regras com menos antecedentes, com menos operadores de concentração/dilatação, negação e indivíduos com menos $f_d(\mu_{A_{j1}}(x_{p1})..., \mu_{A_{jK}}(x_{pK}))$ e, portanto, com uma menor base de regras fuzzy. Com a avaliação, cada indivíduo pode ser selecionado e recombinado para a geração de uma nova população. Este processo transcorre até que o critério de parada seja atingido. Neste instante, a última população é retornada.

4. ESTUDO DE CASOS

4.1. Descrição dos Experimentos

Além do modelo GPFIS, outros modelos para classificação foram também usados. A Tabela 2 apresenta cada um deles, com os parâmetros que os compõem. Cabe ressaltar que a escolha do valor para cada parâmetro deveu-se a testes preliminares efetuados, visando à seleção da melhor configuração. Para o Pitt-GFS foi considerado o mesmo número de funções de pertinência e perfil usados no GPFIS (Figura 4) para cada atributo, de forma a tornar as abordagens mais próximas possíveis.

O poço analisado possui 5995 padrões no total. A forma de avaliar a acurácia de cada método foi a validação cruzada em 10 pastas (10-fold-cv), perfazendo em cada pasta o total de 5394 padrões de treino e 601 de teste. Os resultados relatados são frutos da média de 3 execuções ëm cada pasta do 10-fold-cv para cada método. As métricas calculadas foram a acurácia total, que não discrimina o desbalanceamento entre as classes e a precisão média. A grande vantagem da precisão média para a aplicação compreende o fato de esta levar em conta o desbalanceamento entre as classes (Arenito possui mais padrões do que Background, por exemplo) e, portanto, penalizar classificadores que privilegiam mais a classe dominante em detrimento das demais.

A Tabela 3 apresenta os parâmetros usados no modelo GPFIS.

No caso da PGMG, o processo utilizado foi similar ao de [22], concernente ao tratamento de um problema de múltiplas classes como de classes binárias. Para tanto, a PGMG é executada quatro vezes, dividindo de forma equivalente o número de avaliações factíveis (5000

Tabela 2: Classificadores e parâmetros usados.

Modelo	Parâmetro
GPFIS	Tabela 3
SFGBR do tipo Pittsburgh (Pitt-GFS) [19]	Tabela 3
Bayes Ingênuo (NB) [20]	-
KNN [20]	3-nearest-neighbour, distância euclidiana
Árvore de Classificação (CART) [20]	-
Análise Discriminante Linear (DISC) [21]	-
PGMG	Tabela 3
Perceptron de Múltiplas Camadas (MLP) [20]	Uma camada escondida, função de ativação logística (escondida e saída) e 10 neurônios

Tabela 3: Principais configurações dos modelos baseados em Programação Genética.

Parâmetro	Valor
Tamanho da população	100
Número de gerações	200
Altura máxima da árvore	5
Tamanho do torneio	2
Taxa de cruzamento de alto nível	50%
Taxa de cruzamento de baixo nível	85%
Taxa de mutação	10%
Taxa de clonagem	5%
Taxa de elitismo	1%
Pressão lexicográfica	Sim
Conjuntos Fuzzy de Entrada	Figura 4
Operadores Fuzzy	Produto e Negação

para cada execução), tal que para cada execução é elaborada uma função discriminadora para uma determinada classe. No final das quatro execuções, as funções discriminadoras são reunidas e os padrões separados para a fase de teste são classificados. Para as abordagens evolutivas, a função de avaliação foi o Erro Médio de Classificação.

4.2. Resultados e Discussões

A Tabela 5 exibe os resultados referentes à acurácia dos classificadores em análise. Verifica-se que, em geral, o modelo GPFIS obteve a maior acurácia, em média 4% maior do que o MLP. Em comparação, Pitt-GFS obteve resultados piores do que o GPFIS e o NB. A abordagem padrão - DISC - proporcionou resultados com desempenho pior do que o do GPFIS. A Tabela 6 apresenta a precisão atingida por cada modelo. Ainda, o modelo GPFIS obteve os melhores resultados, proporcionando em média 6,98% maior precisão de classificação do que o NB, o segundo melhor neste quesito.

De forma geral, os resultados do GPFIS são superiores à abordagem do especialista (DISC) em dois quesitos: melhores resultados quando se avalia pelo volume de padrões corretamente classificados (acurácia) e equilíbrio dos esforços para atingir o máximo de padrões de diferentes classes (precisão). Cabe ressaltar que a abordagem NB, que demanda pouco esforço computacional, também auferiu bons resultados e pode ser útil em situações que requeiram aprendizado contínuo e decisões em curtíssimo prazo.

Os resultados indicam que o modelo GPFIS comporta-se relativamente bem em situações

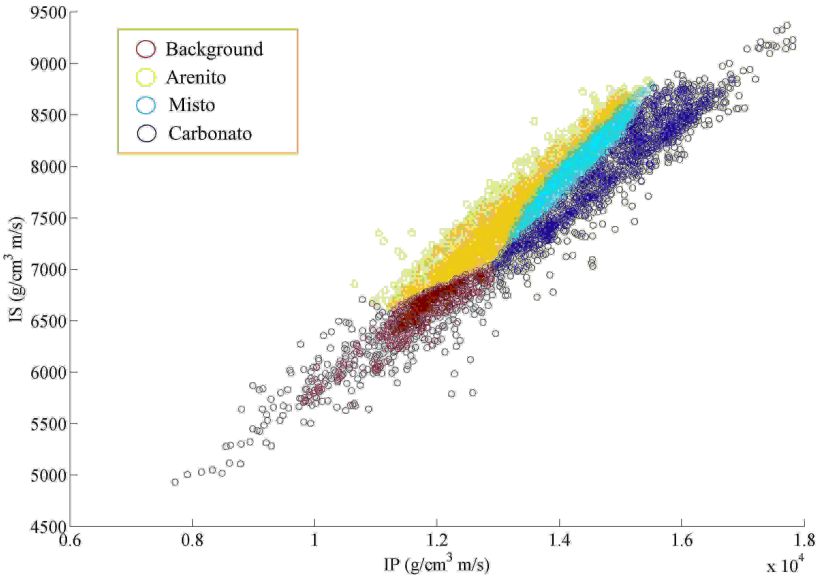

Figura 5: Classes preditas pelo modelo GPFIS.

com baixa intensidade de desbalanceamento entre as classes, quando comparado aos demais classificadores. Uma explicação para isto vem da análise da base de regras fuzzy (Tabela 4) em conjunção com a Figura 5. Para este exemplo, considere-se a litologia Background. Verifica-se que a primeira regra, R1, estabelece que se IS é pequeno (P) então o padrão sísmico pertence ao tipo Background. Ainda, R2 e R3, quando conjuntamente avaliadas, descrevem que se IP é pequeno e IS é pequeno, com IP-IS não sendo nem médio e muito grande, então o padrão sísmico pertence ao tipo de rocha Background. Em suma, pode-se verificar que R1 e R2 delimitam a região em que os padrões da classe Background estão localizados: valores pequenos de IP e IS (Figura 1); R3 coopera com a R2 de modo a se concentrar mais nos padrões que se situam nesta região e relativamente perto de IP e IS médio. A partir dessa construção da região de discriminação, o modelo GPFIS consegue auferir boa classificação para as classes com menos padrões para treinamento.

Tabela 4: Base de regras fuzzy do melhor indivíduo do modelo GPFIS.

Regra	Antecedente	Consequente	Peso
R1	Se IP não é P ou M e IP-IS é G e RP é G	Arenito	0,40
R2	Se RP não é G	Arenito	0,19
R3	Se IP-IS é MG	Arenito	0,41
R4	Se IP é M e RP é G e IP-IS não é P ou M	Misto	0,20
R5	Se IS é G e IP-IS é M ou G ou MG	Misto	0,80
R6	Se IS é MG e IP-IS é M	Carbonato	0,50
R7	Se IP é G e RP é M	Carbonato	0,23
R8	Se IS é G e RP é M	Carbonato	0,52
R9	Se IP-IS não é M e RP não é G	Carbonato	0,18
R10	Se IS é P	Background	0,25
R11	Se IP é P e IP-IS não é M	Background	0,49
R12	Se IP não é MG e IS é P e IP-IS não é MG	Background	0,26

A Figura 5 apresenta a região de discriminação do modelo GPFIS com a base de regras conforme Tabela 4. Observa-se que os padrões da litologia Background se localizam na faixa de valores pequenos e médios de IP e IS. Quando um padrão sísmico possui valores elevados de IP e IS, segundo o modelo GPFIS ele é classificado como Carbonato. Por fim, suponha um padrão sísmico com IP por volta de 14000 $g/cm^3m/s$ e IS de 8200 $g/cm^3m/s$. Após computar IP-IS e RP é possível avaliar que o grau de pertinência desse padrão para cada litologia é: Background=0,00, Carbonato=0,02, Arenito = 0,60 e Misto = 0,398. Um

especialista pode interpretar este resultado de duas maneiras: (i) definir como rocha do tipo Arenito, a partir do critério de decisão da classe mais compatível; (ii) estabelecer que este padrão possui em torno de 60,00% de Arenito, 39,80% de Misto e traços de Carbonato (talvez por medições ruidosas). Esta última é devida à escassez de homogeneidade do padrão sísmico (por exemplo, grandes amostras, ou áreas com diferentes topologias). Ambos os tipos de interpretabilidades podem ser úteis aos especialistas e são viáveis a partir de um Sistema Fuzzy para Classificação como o GPFIS.

Tabela 5: Acurácia média na fase de teste das 3 execuções por pasta de validação cruzada.

Pasta	MLP	Pitt-GFS	GPFIS	PGMG	DISC	NB	CART	KNN
I	69,11%	57,22%	69,17%	47,17%	61,00%	**69,50%**	46,83%	48,33%
II	66,33%	57,15%	**68,73%**	57,48%	57,26%	57,76%	52,25%	53,92%
III	55,87%	54,09%	55,98%	44,69%	44,07%	**60,43%**	44,74%	45,74%
IV	60,60%	59,04%	63,44%	53,20%	46,74%	**69,95%**	45,58%	41,90%
V	51,53%	54,42%	**68,78%**	54,65%	50,25%	68,61%	56,09%	53,26%
VI	**62,99%**	55,04%	61,83%	56,82%	61,10%	51,92%	44,91%	46,91%
VII	46,52%	34,17%	**65,55%**	50,81%	40,57%	35,06%	43,57%	45,58%
VIII	51,78%	**52,06%**	49,83%	47,56%	46,67%	46,50%	41,17%	47,00%
IX	54,44%	54,22%	**58,50%**	58,22%	45,67%	51,33%	41,67%	45,67%
X	**57,79%**	49,81%	55,57%	52,63%	39,43%	46,26%	44,43%	48,25%
Média	57,70%	52,72%	**61,74%**	52,32%	49,28%	55,73%	46,12%	47,66%

Tabela 6: Precisão média na fase de teste das 3 execuções por pasta de validação cruzada.

Pasta	MLP	Pitt-GFS	GPFIS	PGMG	DISC	NB	CART	KNN
I	60,16%	47,95%	**64,10%**	44,66%	61,30%	61,38%	41,35%	44,23%
II	60,05%	54,64%	**66,66%**	49,73%	60,84%	53,63%	47,00%	49,62%
III	56,95%	53,65%	60,05%	38,49%	51,95%	**61,47%**	46,57%	47,90%
IV	58,20%	58,39%	65,39%	46,33%	54,11%	**70,26%**	45,97%	44,07%
V	43,67%	56,44%	**69,66%**	50,23%	55,69%	67,89%	57,84%	56,08%
VI	58,22%	51,65%	61,64%	50,08%	**63,16%**	51,00%	42,31%	45,85%
VII	43,47%	34,39%	**64,47%**	42,79%	46,62%	36,11%	40,45%	41,95%
VIII	49,23%	52,99%	52,21%	43,10%	**52,50%**	46,70%	38,71%	46,11%
IX	48,43%	49,69%	**54,49%**	50,16%	44,82%	47,49%	38,37%	41,22%
X	**50,43%**	44,19%	48,87%	46,94%	36,66%	41,73%	39,20%	43,41%
Média	52,88%	50,40%	**60,75%**	46,25%	52,77%	53,77%	43,78%	46,04%

5. CONCLUSÕES

Este trabalho apresentou uma investigação sobre discriminação litológica a partir de padrões sísmicos. Foram aplicados diversos classificadores, dentre os quais Estatísticos (Bayes Ingênuo, Análise Discriminante, etc.) e por Computação Inteligente (Rede Neural, Classificador Fuzzy-Genético, etc.). O modelo GPFIS proporcionou os melhores resultados, em média, do que os demais modelos. Devido a presença de desbalanceamento das classes, foi providenciado uma explicação do motivo pelo qual o GPFIS obteve bons resultados, a partir da análise da base de regras fuzzy. Por fim, duas interpretações distintas da base de regras fuzzy foram apresentadas. Trabalhos futuros devem explorar técnicas de pré-processamento, por exemplo, métodos de sobre-amostragem para reduzir o efeito do desbalanceamento dos padrões das classes. Outras abordagens, tais como o uso de comitês de classificadores podem auxiliar no desempenho da tarefa de classificação, ou ainda avaliar a metodologia para demais campos exploratórios.

Referências Bibliográficas

[1] Schlumberger, Schlumberger Oilfield Glossary, (http://www.glossary.oilfield.slb.com/ Display.cfm?Term=lithology). Visualizado em Março de 2014.

[2] U.S. Geological Survey, Earthquake Glossary, (http://earthquake.usgs.gov/learn/ glossary/?term=lithology). Visualizado em Março de 2014.
[3] Doveton, J.H. The Geological Application of Wireline Logs: A Keynote Perspective. *AAPG Methods in Exploration*, v. 13, p. 115–122, 2002.
[4] Busch, J.M.; Fortney, W.G.; Berry, L.N. Determination of lithology from well logs by statistical analysis: *Society of Petroleum Engineers Formation Evaluation*, v. 2, p. 412–418, 1987.
[5] Rogers, S.J.; Fang, J.H.; Karr, C.L.; Stanley, D.A. Determination of lithology from well logs using a neural network. *AAPG Bulletin*, v. 76, n. 5, p. 731-739, 1992.
[6] Santos, R.O.V.; Vellasco, M.M.B.R.; Artola, F.A.V.; Da Fontoura, S.A.B. Neural Net Ensembles for Lithology Recognition. In: Windeatt, T.; Roli, F. (eds.), *Multiple Classifier Systems*, volume 2709, Lecture Notes in Computer Science, p. 246–255. Springer: Heidelberg, 2003.
[7] Leite, V.R.C. Uma análise da classificacão de litologias utilizando SVM, MLP e métodos Ensemble. Dissertacão de Mestrado. Departamento de Informática. Rio de Janeiro: Pontifícia Universidade Católica do Rio de Janeiro, 2012.
[8] Saggaf, M.M.; Nebrija E.L. A fuzzy logic approach for the estimation of facies from wire-line logs. *AAPG Bulletin*, v. 87, n. 7, p. 1223–1240, 2003.
[9] Córdon, O.; Herrera, F.; Hoffmann, F.; Magdalena, L. *Genetic Fuzzy Systems. Evolutionary Tuning and Learning of Fuzzy Knowledge Bases*. World Scientific, 2001.
[10] Herrera, F. Genetic Fuzzy Systems: taxonomy, current research trends and prospects. *Evolutionary Intelligence*, v.1 ,n.1 ,p.27-46, 2008.
[11] Koshiyama, A.S.; Vellasco, M.M.B.R.; Tanscheit, R. GPFIS: Um Sistema Fuzzy-Genético Genérico baseado em Programação Genética. Dissertação de Mestrado (em fase de publicação online). Departamento de Engenharia Elétrica. Rio de Janeiro: Pontifícia Universidade Católica do Rio de Janeiro, 2014, p.225.
[12] Avseth, P.; Mukerji, T.; Mavko, G. *Quantitative Seismic Interpretation: Applying Rock Physics Tools to Reduce Interpretation Risk*. Cambridge: Cambridge University Press, 2005.
[13] Han, D. Effects of porosity and clay content on acoustic properties of sandstone and unconsolidated sediments. Unpublished Ph.D. dissertation, Stanford University, 1986.
[14] Telford, W.M.; Geldart, L.P.; Sheriff, R.E. *Applied Geophysics*. Cambridge: Cambridge University Press, 1990.
[15] Castagna, J.P.; Batzle, M.L.; Kan, T.K. Rock physics: The link between rock properties and AVO response. *Investigations in Geophysics*, v. 8, p.135-171, 1993.
[16] Searson, D.; Willis, M.; Montague, G. Coevolution of nonlinear PLS model components. *Journal of Chemometrics*, v.21, n. 12, p.592-603, 2007.
[17] Hinchliffe, M.; Hiden, H.; McKay, B.; Willis, M.; Tham, M.; Barton, G. Modelling Chemical Process Systems Using a Multi-Gene. Late Breaking Papers at the Genetic Programming, p. 56-65, 1996.
[18] Poli, R.; Langdon, W.B.;, McPhee, N.F.; Koza, J.R. A field guide to genetic programming. Rayleigh: Lulu, 2008.
[19] Bastian, A. Identifying fuzzy models utilizing genetic programming. *Fuzzy Sets and Systems*, v.113, n.3, p.333-350, 2000.
[20] Mitchell, T.M. *Machine learning*. Burr Ridge: McGraw Hill, 1997.
[21] Johnson, R. A.; Wichern, D. W. *Applied multivariate statistical analysis*. 5 ed. New Jersey: Prentice-Hall, 2002.
[22] Kishore, J. K. ; Patnaik, L. M. ; Mani, V.; Agrawal, V. K. Application of genetic programming for multicategory pattern classification. *IEEE Trans. Evol. Comput.*, v.4, n.3, p. 242–258, 2000.