Lithofacies classification using Optimum-Path Forest method

Ivan Mingireanov Filho, Alexandre Campane Vidal.
State University of Campinas - UNICAMP

Abstract

Pattern classification aims to find the label of each sample using a feature vector in dataset samples. As traditional methods usually does, the learning process distributes samples into different labels using a training subset and determines rules to decide the classification of each subset. The pattern recognition algorithm, based in Optimum-Path Forest (OPF) used here, models the problem like a graph. In this graph the nodes are the samples and the arcs are defined by some adjacency relation, the most relevant samples are called prototypes. A competition process among samples starts offering optimum-path costs to the remaining dataset samples. Using the complete graph in adjacency relation the competition becomes global, not based only in local geometry as traditional methods. The search is done finding the prototypes samples that belongs to overlap regions and boundaries between the labels in the training set. These regions are very susceptible to misclassification. The prototypes samples offer optimum-path cost computed as the maximum path arc-weight between these prototypes and the other dataset samples. These arc-weights are computed by the distance between their features vectors. The goal of this work is validate the OPF method classifying non-cored wells information set, from *Campo de Namorado*, in *Bacia de Campos*, RJ, freely available by National Petroleum Agency (ANP). The dataset is composed by 5 attributes for each depth, totaling 4732 samples, which 1950 are core samples. The results show the time spent on both, training and classification process. They were very short always less than 1 second. In cross-validation, 87,1% of the core samples maintained their label. The average accuracy computed to each well was 64,8%, and using all core samples in training process the accuracy was 76,8%. These results suggest a new approach suitable to process a large volume dataset in a short time, being an alternative for non-cored wells classification.

Introdução:

O método de classificação por Floresta de Caminhos Ótimos (OPF) utiliza a Transforma Imagem Floresta (IFT), uma ferramenta geral para modelar, implementar e avaliar operadores baseados em conexidade (Falcão *et al.,* 2004), reduzindo o problema de classificação ao cálculo de uma floresta de caminhos ótimos em um grafo derivado do conjunto de dados das amostras. O valor de um caminho é normalmente calculado por uma função dependente dos atributos e da posição do atributo ao longo do caminho do grafo (Papa, 2009).

Para utilização do classificador baseado em OPF com aprendizado supervisionado, o modelamento do problema de reconhecimento de padrões é considerado como uma problema de OPF em um grafo definido no espaço de atributos, onde os nós são as amostras, as quais são representadas pelos seus respectivos vetores da atributos, e os arcos são definidos de acordo com uma relação de adjacência pré-estabelecida. Tanto os nós quanto os arcos podem ser ponderados por alguma métrica de distância aplicada a seus vetores de atributos, diferente dos métodos tradicionais por não utilizar a ideia de geometria do espaço de atributos, o que permite melhores resultados em bases com *outliers* e sobreposição de classes. Diversas funções de custo podem ser empregadas para particionar o grafo em árvores de caminhos ótimos, as quais são enraizadas pelos seus respectivos protótipos (sementes) na fase de treinamento. O rótulo de uma amostra a ser classificada é o mesmo do protótipo mais fortemente conexo a ela. A abordagem adotada utiliza como relação de adjacência ao grafo completo e de busca, durante a fase de treinamento, os elementos mais representativos de cada classe como sendo elementos que pertençam à intersecção entre as classes no conjunto de treinamento (Papa, 2009).

Os protótipos participam de um processo de competição disputando as outras amostras oferecendo-lhes caminhos de menor custo e seus respectivos rótulos. Ao final deste processo, obtêm-se um conjunto de treinamento particionado em árvores de caminhos ótimos, sendo que a união das mesmas nos remete a uma floresta de caminhos ótimos. Esta abordagem apresenta vários benefícios com relação a outros métodos de classificação de padrões supervisionados: (i) é livre de parâmetros, (ii) possui tratamento nativo de problemas multiclasses e (iii) não faz alusão sobre forma e/ou separabilidade das classes (Papa, 2009).

O algoritmo OPF com grafo completo tem sido amplamente utilizado em diversas aplicações, tais como avaliação de descritores de textura, diagnóstico automático de patologias na laringe e classificação de impressões digitais (Papa, 2009). Essas aplicações fundamentam a aplicação em dados de poços para classificação da litologia.

A fase de treinamento do classificador baseado em OPF usando o grafo completo consiste em encontrar o conjunto S de protótipos, ou seja, os elementos mais representativos de cada classe. Essa escolha pode ser de várias maneiras heurísticas, inclusive, por uma escolha aleatória de protótipos. Entretanto, tal escolha pode prejudicar o desempenho do classificador, tornando-o instável e com um alto grau de sensibilidade com relação aos protótipos escolhidos. Deseja-se estimar os protótipos nas regiões de sobreposição de amostras e nas fronteiras entre as classes, visto que são regiões muito susceptíveis a erros de classificação. Outra técnica que pode ser abordada é a utilização do grafo k-vizinhos mais próximos (k-NN), onde estima estes protótipos nos pontos de alta concentração de amostras (Papa, 2009).

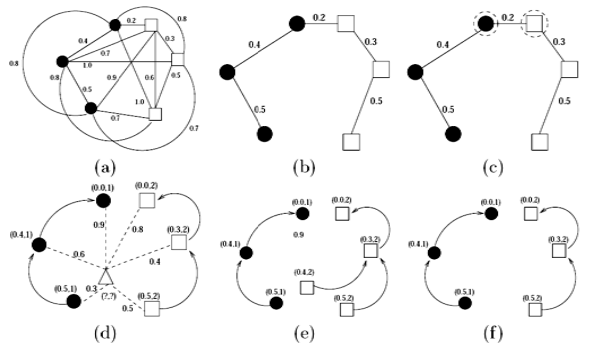

Fig. 1: *(a) Grafo completo ponderado nas arestas para um determinado conjunto de treinamento. (b) Árvore Geradora Mínina (MST) do grafo completo. (c) Protótipos escolhidos como sendo os elementos adjacentes de classes diferentes na MST (nós circulados). (d) Floresta de caminhos ótimos resultante para a função de valor de caminho f_{max} e dois protótipos. Os identificadores (x, y) acima dos nós são, respectivamente, o custo e o rótulo dos mesmos. A seta indica o nó predecessor no caminho ótimo. (e) Uma amostra de teste (triângulo) da classe 2 e suas conexões (linhas pontilhadas) com os nós do conjunto de treinamento. (f) O caminho ótimo do protótipo mais fortemente conexo, seu rótulo 2 e o custo de classificação 0.4 são associados à amostra de teste. Note que, mesmo a amostra de teste estando mais próxima de um nó da classe 1, ela foi classificada como sendo da classe 2 (Papa, 2009).*

Computando uma árvore mínima geradora (MST) no grafo completo (Z_1, A) na figura 1(a), obtemos um grafo conexo acíclico cujos nós são todas as amostras em Z_1, e os arcos são não direcionados e ponderados, figura 1(b). Seus pesos são dados pela distância d entre os vetores de atributos de amostras adjacentes. Esta árvore de espalhamento é ótima no sentido em que a soma dos pesos de seus arcos é mínima se comparada a outras árvores de espalhamento no grafo completo. Os protótipos a serem escolhidos são os elementos conectados na MST com diferentes rótulos em Z_1, ou seja, elementos mais próximos de classes diferentes, figura 1(c). Removendo-se os arcos entre classes diferentes, tais amostras adjacentes tornam-se protótipos em S e pode-se computar uma floresta de caminhos ótimos em Z_1, figura 1(d). Nota-se que uma dada classe pode ser representada por múltiplos protótipos e deve existir pelo menos um protótipo por classe. A ideia consiste em ponderar os arcos entre amostras de diferentes classes com um valor muito alto, impossibilitando assim que os protótipos de uma classe conquistem elementos de outras classes (Papa, 2009).

Para a classificação, todos os arcos são considerados conectando a amostra teste com as amostras do grafo original, onde na figura 1 (d) está representado pelo triângulo. Considerando todos esses possíveis caminhos, desejamos encontrar o caminho ótimo com a classe de seu protótipo mais fortemente conexo. Assim, a classificação simplesmente associa a amostra à classe que satisfaz essas condições, figura 1 (e). Vale ressaltar que, embora a amostra a ser classificada esteja mais próxima de um elemento de classe bola, figura 1 (c), a mesma é classificada como sendo da classe quadrado, o que demonstra que os classificadores baseados em OPF utilizam a força de conexidade entre as amostras para a classificação dos dados, ou seja, não são algoritmos baseados em conexidade local apenas como, por exemplo, os classificadores de redes neurais (NN) e k-vizinhos próximos (k-NN) (Papa, 2009).

O classificador baseado em OPF com grafo completo assemelha-se ao algoritmo vizinho mais próximo (NN) somente no caso onde todos os elementos do conjunto de treinamento são considerados protótipos, sendo este um caso atípico onde, certamente, o conjunto de atributos escolhido não foi o mais adequado para a representação do conjunto de dados. Outra diferença a ser considerada é que os classificadores NN e k-vizinhos mais próximos (k-NN) tomam uma decisão local para a classificação dos dados, ao contrário dos classificadores baseados em OPF, os quais possibilitam uma solução em âmbito global, criando uma floresta de caminhos ótimos que mapeia, para cada amostra do conjunto de dados, o caminho ótimo entre ela e o seu protótipo mais fortemente conexo (Papa, 2009).

O objetivo do presente trabalho foi classificar um conjunto de informações de poços não-testemunhados utilizando o método de classificação supervisionada de padrões por floresta de caminhos ótimos, com validação cruzada.

Metodologia

Foram utilizados um conjunto de informações de diferentes atributos referentes ao Campo de Namorado, da Bacia de Campos, RJ, disponibilizado livremente pela Agência Nacional de Petróleo (ANP). Dentre todas informações fornecidas, foram utilizados 5 atributos, de acordo com a tabela 1.

Tabela 1: Atributos utilizados para implementação do classificador baseado em OPF.

Atributos Avaliados	Sigla	Aplicação
Densidade	RHOB	Detectação dos raios gamas defletidos pelos elétrons por uma fonte situada dentro de um poço
Porosidade Neutrônica	NPHI	Medida da quantidade de nêutrons da rocha após o bombardeio
Sônico	DT	Medida da diferença dos tempo de trânsito de uma onda mecânica através das rochas
Indução	ILD	Leitura aproximada da resistividade
Raio Gama	GR	Detecção da radioatividade total da formação geológica

O conjunto de dados contém informações medidas dos atributos de 7 poços, em função da profundidade, no total de 4732 amostras, das quais 1950 possuem testemunho, ou seja 41,2% do total. É um conjunto privilegiado pela quantidade de testemunhos dos poços, pois normalmente esse número é bem menor, não ultrapassando os 20% do total, devido ao alto custo e dificuldade de gerar os testemunhos.

Para a classificação, de acordo com o interesse do trabalho proposto, agrupou-se as fácies em quatro classes, de acordo com a tabela 2. São essas rochas sem testemunho o objetivo de classificação pelo o método OPF, totalizando de 2782 amostras.

Tabela 2: Descrição das classes para classificação das rochas do Campo de Namorado pelo método OPF.

Classe	Descrição
1	Rochas Reservatório
2	Rochas Possível Reservatório
3	Rochas Não Reservatório
4	Rochas Sem Testemunho

As informações foram separadas para cada um dos 7 poços, de acordo com a tabela 3, e foram realizadas duas baterias de classificação após o treinamento do algoritmo, utilizando o algoritmo 3 - *Classificador Supervisionado baseado em Floresta de Caminhos Ótimos usando grafo completo*, disponibilizado em linguagem C para utilização em plataforma Linux em www.ic.unicamp.br/~afalcao/LibOPF (Papa, 2009).

Tabela 3: Distribuição das classes por poço.

	Bateria 1					*Bateria 2*
Poço	Mantiveram		Alteraram		Acurácia	Acurácia
1	23	60,5%	15	39,5%	59,0%	75,1%
2	49	92,4%	4	7,6%	69,7%	75,7%
4	138	85,7%	23	14,3%	71,9%	78,4%
7	145	80,5%	35	19,5%	66,5%	77,3%
11	262	93,6%	18	6,4%	67,6%	85,3%
234	192	96,9%	6	3,1%	53,6%	75,9%
42	53	84,1%	10	15,9%	65,6%	68,6%

A primeira bateria, foi feita apenas com as informações de cada poço, individualmente, utilizando metade dos dados de rochas com testemunhos do poço para treinar o classificador. Na sequência, foram classificados os dados sem testemunhos juntos com a outra metade de dados com testemunhos, para validação cruzada.

A segunda bateria agrupou todos os 1950 dados de rochas com testemunhos para treinamento, e classificou os dados sem testemunhos e metade dos dados com testemunhos, individualmente para cada um dos 7 poços. Ou seja, a diferença entre as duas baterias de classificação foi na fase de treinamento, onde a segunda utiliza todo o conjunto de informações dos testemunhos, com o intuito de melhorar esta fase pela quantidade de informações maior, mantendo o conjunto de classificação e de validação cruzada.

Resultados

Os resultados obtidos foram bastante satisfatórios em ambas as baterias. Na bateria 1, a classificação das amostras sem testemunhos foi feita após treinamento do algoritmo com metade das amostras com testemunhos, e a outra metade agrupada com as amostras sem testemunhos. Por essa validação cruzada, algumas dessas amostras com testemunhos foram reclassificadas, alterando a classe original, e outras se mantiveram. Os números absolutos e percentuais das amostras que mantiveram a classe e das amostras que foram alteradas de classe estão apresentados a seguir na tabela 4.

Tabela 4: Resultados obtidos das duas baterias de classificação

Poço	*Classe 1*	*Classe 2*	*Classe 3*	*Classe 4*	*Número Total de Amostras*
1	4,76%	3,13%	1,75%	90,36%	799
2	9,26%	0,93%	6,17%	83,64%	648
4	25,65%	5,34%	24,61%	44,41%	581
7	10,59%	6,98%	29,07%	53,36%	774
11	6,49%	1,40%	90,53%	1,58%	570
234	1,47%	0,00%	51,54%	46,99%	747
42	16,31%	277%	1,47%	79,45%	613

Ainda nessa bateria, são apresentadas as medidas de acurácia, que o próprio algoritmo determina. A acurácia utiliza em sua função um erro calculado quando amostra a ser rotulada não recebe o rótulo determinado corretamente no treinamento. Quanto menor for este erro, maior será a acurácia.

Já na bateria 2, foram utilizadas todas as 1950 amostras com testemunhos de todos os poços na fase de treinamento, para realizar um treino com um banco de dados maior. A classificação foi realizada como na bateria 1, isto é, os dados a serem classificados eram as amostras sem testemunhos agrupados com a outra metade das amostras com testemunhos. Por isso, ao analisar a validação cruzada, não foi detectada nenhuma alteração nas amostras com testemunhos, pois estas amostras já estavam contidas no conjunto de treinamento. Mas, como esperado, ao utilizar um conjunto maior de treinamento, a tendência do elemento

mais representativo da classe ser escolhido durante a competição é maior, diminuindo o erro, aumentando significativamente a acurácia.

Conclusões

O método de classificação supervisionada de padrões baseada em Floresta de Caminhos Ótimos apresentado em Papa (2009), comprova sua eficiência e confiabilidade em relação à outros métodos de classificação baseados em Máquinas de Vetores de Suporte (SVM), Redes Neurais Artificiais com Perceptrons em Múltiplas Camadas (ANN-MLP), k-vizinhos mais próximos (k-NN) e Classificador Bayesiano (BC).

Esta classificação, realizada em dados de poços, também mostrou-se um método simples, rápido e eficiente. Na bateria 1, houve a manutenção de 87,1% das classes originais na validação cruzada, e acurácia média de cerca de 64,8%. Na bateria 2, utilizando todo o conjunto de amostras com testemunhos, a acurácia média sobe para cerca de 76,8%.

O método apresenta também algumas qualidades interessantes, como não ter a necessidade de que o número de classes das amostras sejam iguais. Corroborando com essa observação, os dados utilizados aqui neste trabalho, em alguns poços, não foram observados amostras de uma certa classe, ou em alguns casos, possuíam um número pequeno para certa classe.

Em Papa 2009 é apresentado também uma outra abordagem, utilizando o grafo k-vizinhos mais próximos (k-NN), onde a diferença básica é que a abordagem aqui utilizada faz uso do grafo completo para estimar protótipos na fronteira das classe, enquanto a outra estima estes protótipos nos pontos de alta concentração de amostras.

Agradecimentos

Gostaríamos de agradecer à Agência Nacional de Petróleo (ANP) pela disponibilização dos dados para estudos e pelo apoio financeiro.

Referências

Hecht-Nielsen R. (Eds.) 1990. Neurocomputing. Addison-Wesley Publishing Company, Redwood City, 433 pp.

Bishop C.M. 1994. Neural networks and their applications, Neural Computing Reseach Group, United Kingdom.Rev. Sci. Instrum. **65 (6):** 1803-1832.

Papa J.P. 2009. Classificação supervisionada de padrões utilizando floresta de caminhos ótimos. Tese de Doutorado, Instituto de Computação, Unicamp, 75 p.

Falcão A.X., Stolfi J., Lotufo R.A. 2004. The image foresting transform: theory, algorithms, and applications. IEEE transactions on pattern analysis and machine intelligence.**26 (1):** 19-29.