

Copyright 2005, Instituto Brasileiro de Petróleo e Gás - IBP

Este Trabalho Técnico Científico foi preparado para apresentação no 3º Congresso Brasileiro de P&D em Petróleo e Gás, a ser realizado no período de 2 a 5 de outubro de 2005, em Salvador. Este Trabalho Técnico Científico foi selecionado e/ou revisado pela Comissão Científica, para apresentação no Evento. O conteúdo do Trabalho, como apresentado, não foi revisado pelo IBP. Os organizadores não irão traduzir ou corrigir os textos recebidos. O material conforme, apresentado, não necessariamente reflete as opiniões do Instituto Brasileiro de Petróleo e Gás, Sócios e Representantes. É de conhecimento e aprovação do(s) autor(es) que este Trabalho será publicado nos Anais do 3º Congresso Brasileiro de P&D em Petróleo e Gás

COMPARAÇÃO ENTRE MÉTODOS DE AGRUPAMENTOS DE DADOS EM ATRIBUTOS SÍSMICOS

Daniel R. S. Moraes¹, Alexandre G. Evsukoff², Nelson F. F. Ebecken³

¹ COPPE/UFRJ, CP. 68506, CEP. 21941-972, Rio de Janeiro - RJ, dsilos@gmail.com

² COPPE/UFRJ, CP. 68506, CEP. 21941-972, Rio de Janeiro - RJ, evsukoff@coc.ufrj.br

³ COPPE/UFRJ, CP. 68506, CEP. 21941-972, Rio de Janeiro - RJ, nelson@ntt.ufrj.br

Resumo - Métodos de agrupamento de dados são ferramentas importantes em mineração de dados, pois permitem detectar padrões não encontrados explicitamente nos dados armazenados. Neste trabalho, métodos de agrupamento são aplicados a atributos sísmicos visando detectar padrões que permitam identificar diferentes litologias na análise exploratória de petróleo.

Atributos sísmicos são gerados a partir de transformações sobre os dados de uma linha sísmica, sendo alguns selecionados para formar um conjunto de dados a ser aplicado pelos diferentes algoritmos de agrupamento de dados.

Este trabalho apresenta a utilização de três métodos de agrupamento de dados diferentes: As redes neurais conhecidas por SOM (*Self-Organizing Maps*) e UVQ (*Unsupervised Vector Quantiser*) e o algoritmo baseado em teoria de conjuntos nebulosos conhecido por FCM (*Fuzzy C-Means*).

Na aplicação de cada método, várias partições são geradas e é aplicada uma função de validação de agrupamento de dados conhecida por índice PBM. Esta função indica o número do particionamento em que os dados foram melhor agrupados e os resultados são apresentados e analisados.

Palavras-Chave: Agrupamento de Dados; Redes Neurais; Conjuntos Nebulosos, Validação de Agupamentos de Dados.

Abstract – Clustering methods are important tools in data mining, allowing the detection of hidden patterns that are not explicitly recorded in data. Their applications to seismic attributes have the aim of detecting seismic patterns that would enable the identification of lithologies in exploration analysis of Petroleum.

Seismic attributes are generated from transformation of the data from a seismic line. Some attributes are selected in order to form a data set that is used by the clustering methods.

This work presents an application using three distinct clustering methods: artificial neural networks known as SOM (*Self-Organizing Maps*) and UVQ (*Unsupervised Vector Quantiser*) and the algorithm based on fuzzy sets known as FCM (*Fuzzy C-Means*).

In the application of each method, several partitions are generated and a cluster validity function, called PBM index, is applied. This function shows the partition number corresponding to the best clustering and the results are presented and analyzed.

Keywords: Clustering, Neural Networks, Fuzzy Sets, Cluster Validity.

1. Introdução

O uso de algoritmos de Inteligência computacional como Redes Neurais Artificiais, Lógica Nebulosa (*"Fuzzy Logic"*), Algoritmos Genéticos e algoritmos híbridos vêm sendo utilizados em caracterização de reservatórios, classificação e previsão de porosidade, classificação não-supervisionada entre dados sísmicos e perfis de poços, etc.

Neste trabalho, algoritmos são utilizados em uma abordagem de aprendizado não-supervisionado objetivando-se agrupar os dados (atributos sísmicos) em diferentes agrupamentos (clusters). Dois modelos de redes neurais e um

modelo de lógica nebulosa são utilizados. Os dados são separados por similaridade entre eles com o objetivo de identificar padrões sísmicos que separem diferentes litologias.

O Objetivo deste trabalho é analisar o desempenho de métodos de agrupamento de dados sobre uma base de dados composta por diferentes atributos sísmicos. Esta avaliação é realizada através da aplicação do índice de validação de agrupamento de dados conhecido por PBM, cujo nome é formado pelas iniciais dos sobrenomes dos autores (Pakhira, Bandyopadhyay e Maulik). O índice é aplicado aos resultados dos diferentes algoritmos de agrupamentos, sendo que em cada algoritmo é determinado um número inicial e um número final de agrupamentos. O maior índice encontrado corresponde à melhor configuração que é composta pelo algoritmo de agrupamento e um determinado número de particionamentos.

Este artigo está organizado da seguinte forma: A próxima seção apresenta os Métodos de Agrupamentos de Dados utilizados na aplicação; a seção 3 descreve o índice PBM, utilizado para validar os resultados gerados pelos métodos de Agrupamentos de Dados; a seção 4 apresenta detalhes sobre a base de dados e os aplicativos utilizados. A seção 5 descreve os resultados gerados pela aplicação do índice PBM; a seção 6 apresenta a conclusão deste trabalho.

2. Métodos de Agrupamentos de Dados

Agrupamento de Dados é uma das tarefas mais usuais em processos de mineração de dados na descoberta e identificação de distribuições e padrões de interesse. Através da aplicação de um método de agrupamento de dados é possível agrupar um conjunto de dados em classes de elementos similares. Estes métodos são caracterizados pela aplicação de algoritmos de classificação não-supervisionada, cujo objetivo é separar um conjunto de dados utilizando-se apenas informações provenientes dos atributos que compõem o conjunto de dados.

Um método de agrupamento de dados separa um conjunto de dados em um determinado número de grupos, definido pelo usuário. Tais grupos são formados pela similaridade existente entre os próprios dados de entrada. Eles podem representar a existência de padrões não naturalmente detectáveis, isto é, informações e conhecimentos existentes e que permanecem ocultos na maneira tradicional em que os dados são armazenados podem ser descobertos aplicando-se técnicas de agrupamento de dados. Este trabalho apresenta métodos baseados em redes neurais (redes UVQ e SOM) e conjuntos nebulosos (algoritmo FCM).

Considere um conjunto de dados $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ contendo n registros do vetor de observações ou atributos

$\mathbf{x}$. Uma partição de um conjunto em grupos C_i satisfaz às seguintes condições: $u_{ti} = 1$ se $\bar{\mathbf{x}}_t \in C_i$, caso contrário $u_{ti} = 0$. O propósito é agrupar um conjunto de dados tal que $C_i \neq \emptyset$; para $i = 1, \dots, K$, $C_i \cap C_j = \emptyset$; para $i = 1, \dots, K$, $j = 1, \dots, K$ e $i \neq j$.

A matriz de partição $\mathbf{U}(\mathbf{X})$ de tamanho $n \times K$, define os agrupamentos, através de seus elementos u_{ti} , que representa o valor de pertinência de $\mathbf{x}_t$, ao agrupamento C_i . Desta forma, $u_{ti} = 1$ indica que o registro t pertence ao grupo C_i e $u_{ti} = 0$, caso contrário.

2.1. Redes Neurais

Redes Neurais Artificiais ou simplesmente, Redes Neurais são modelos de computação caracterizada por sistemas que relembram (em algum nível) a estrutura do cérebro humano. De acordo com Haykin (2001): "Uma rede neural é um processador paralelamente distribuído constituído de unidades de processamento simples, que tem a propensão natural de armazenar conhecimento experimental e torná-lo disponível para o uso. Sua semelhança com o cérebro se deve por adquirir conhecimento por meio de processos de aprendizado e pelas "forças" de interconexão entre os neurônios, conhecidas como pesos sinápticos e que são usadas para armazenar o conhecimento adquirido". O uso de redes neurais como ferramentas de agrupamento de dados tem tido expressivo sucesso na Indústria de Petróleo (Fleck et al., 2003; Nikraves and Hassib, 2001; Nikraves et al., 2003).

2.2. Redes Neurais Baseadas em Competição

A forma de competição mais conhecida é a chamada "O vencedor leva tudo" (Braga et al., 2000; Haykin, 2001; Azevedo et al., 2000; Fausett, 1994), onde um grupo de neurônios competem entre si e somente um será o vencedor.

Durante o treinamento, a rede determina a unidade de saída (neurônio vencedor). O vetor de pesos (sinapses) é então ajustado de acordo com um algoritmo de aprendizado da rede que geralmente é a regra ou aprendizado de Kohonen. Nesta forma de aprendizado, o vetor de pesos do neurônio vencedor terá seus pesos atualizados, isto é, será formado um novo vetor de pesos que é uma combinação linear do vetor de pesos antigo e o vetor de pesos atual (Fausett, 1994).

A rede SOM, cuja sigla significa Mapas Auto-Organizáveis (do inglês *Self-Organizing Maps*) é o modelo de redes neurais mais famoso, dentre os propostos por Teuvo Kohonen (Kohonen, 2001). Sua principal característica está no conceito de vizinhança. Ela aprende a reconhecer seções de vizinhanças no espaço de entrada e também a topologia dos vetores de entrada sobre os quais ocorre o treinamento.

Os neurônios na camada de competição de uma rede SOM são arranjados originalmente em posições físicas de acordo com a função de topologia. Pode-se arranjar os neurônios em topologias em forma de grade (retangular) ou hexagonal. Todos os neurônios a uma certa vizinhança do neurônio vencedor são atualizados utilizando-se a regra de Kohonen.

A rede UVQ, Quantização Vetorial Não-Supervisionada (do inglês "*Unsupervised Vector Quantiser*"), é uma versão modificada da rede LVQ ("*Learning Vector Quantiser*") (Braga et al., 2000; Azevedo et al., 2000)

Diferentemente da rede SOM a UVQ possui duas saídas. A primeira refere-se ao número do agrupamento enquanto a segunda representa o grau de competição ("*match degree*"), isto é, uma medida de confiança de quão próximo o vetor de entrada encontra-se de seu centro de classe. A proximidade ao valor 1 indica uma alta proximidade com seu centro de classe. Tal saída pode ser utilizada como um critério indicando a qualidade do agrupamento dos dados pela rede.

2.3. Algoritmo FCM

O Algoritmo FCM (do inglês "*Fuzzy C-Means*") é um dos métodos de agrupamento de dados baseado em conjuntos nebulosos. Ele permite que um dado pertença a dois ou mais agrupamentos simultaneamente (Leski, 2003; Besdek, 1981).

No caso de agrupamento nebulosos, a matriz de partição $\mathbf{U} = [u_{ti}]_{n \times K}$, onde $u_{ti} \in [0,1]$, tal que u_{ti} denote a grade de pertinência do t -ésimo elemento ao i -ésimo agrupamento. As equações 1 e 2 exibem condições que ocorrem em particionamentos nebulosos.

$$0 < \sum_{t=1}^n u_{ti} < n \quad \text{para } i = 1, \dots, K, \quad (1)$$

$$\sum_{i=1}^K \sum_{t=1}^n u_{ti} = n. \quad (2)$$

3. Índice de Validação PBM

Validação de agrupamentos de dados é uma avaliação de partições que objetiva estabelecer qual partição possui uma melhor estrutura de agrupamento em um conjunto de dados não conhecido.

O Índice PBM (Pakhira et al., 2004), foi o índice de validação de agrupamentos de dados utilizado neste trabalho para validar os agrupamentos gerados pelos agrupamentos de dados dos algoritmos descritos na seção anterior.

Para uma melhor compreensão de como o índice PBM é calculado, são apresentadas as seguintes notações (equações 3,4,5 e 6):

O índice PBM é definido nas equações como:

$$PBM(K) = \left(\frac{1}{k} \times \frac{E_1}{E_K} \times D_K \right)^2, \quad (3)$$

onde K é o número de agrupamentos e

$$E_K = \sum_{i=1}^K E_i, \quad (4)$$

$$E_K = \sum_{t=1}^n u_{ti} d(\mathbf{x}_t, \mathbf{w}_i), \text{ tal que } \quad (5)$$

$$D_K = \max_{i,j=1}^K d(\mathbf{w}_i, \mathbf{w}_j). \quad (6)$$

n é o número total de pontos (registros) em um conjunto de dados, $U(X) = [u_{ij}]n \times K$ é uma matriz de partição para os dados e w_i é o centro do i -ésimo agrupamento. O objetivo é maximizar este índice a fim de se obter o número real de agrupamentos, ou seja, o valor máximo deste índice indica o melhor particionamento (Pakhira et al., 2004).

4. Aplicação

A Base de dados utilizada na aplicação contém uma linha sísmica, no formato SEG Y. Os dados foram cedidos à Universidade Federal do Rio de Janeiro pela ANP (Agência Nacional de Petróleo) e originam-se da Bacia de Almada, localizada no litoral do Estado da Bahia. Foi escolhida uma janela sísmica em torno da perfuração de um poço e foram gerados quatro atributos sísmicos: Amplitude Instantânea, Cosseno da Fase, Frequência Média e Energia (os dois últimos a cada 4 ms).

Foram utilizados os seguintes programas: OpendTect, Matlab, SOM PAK e um conjunto de rotinas para Matlab conhecido por SEG Y-MAT, indispensável para importar dados no formato SEG Y no programa Matlab. A figura 1 exibe o esquema da aplicação informando quais programas foram utilizados em quais etapas.

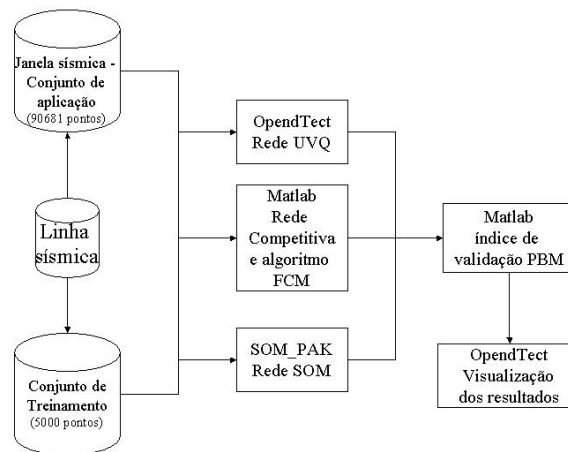


Figura 1 - Esquema da Aplicação

5. Resultados

A tabela 1 mostra os resultados da aplicação do índice PBM aos agrupamentos de dados utilizados e a figura 2 exibe gráficos referentes a estes resultados.

Tabela 1 - Resultados da aplicação do índice PBM

Classes	FCM	SOM (Hexagonal)	SOM. (Retangular)	UVQ
5	3,13E+14	4,33E+14	4,33E+14	2,94E+14
6	3,54E+14	3,63E+14	3,58E+14	1,30E+14
7	2,90E+14	4,34E+14	7,01E+14	3,19E+14
8	2,50E+14	7,21E+14	2,24E+14	8,40E+14
9	6,23E+14	4,63E+15	7,67E+14	4,26E+14
10	6,65E+14	5,29E+15	3,90E+15	4,07E+14
11	6,59E+14	1,94E+15	5,60E+14	3,85E+14
12	6,27E+14	2,65E+15	1,20E+15	3,17E+14
13	5,69E+14	8,42E+14	2,94E+15	2,89E+14
14	5,07E+14	2,61E+15	4,71E+14	2,19E+14
15	4,78E+14	2,41E+15	2,91E+15	2,25E+14

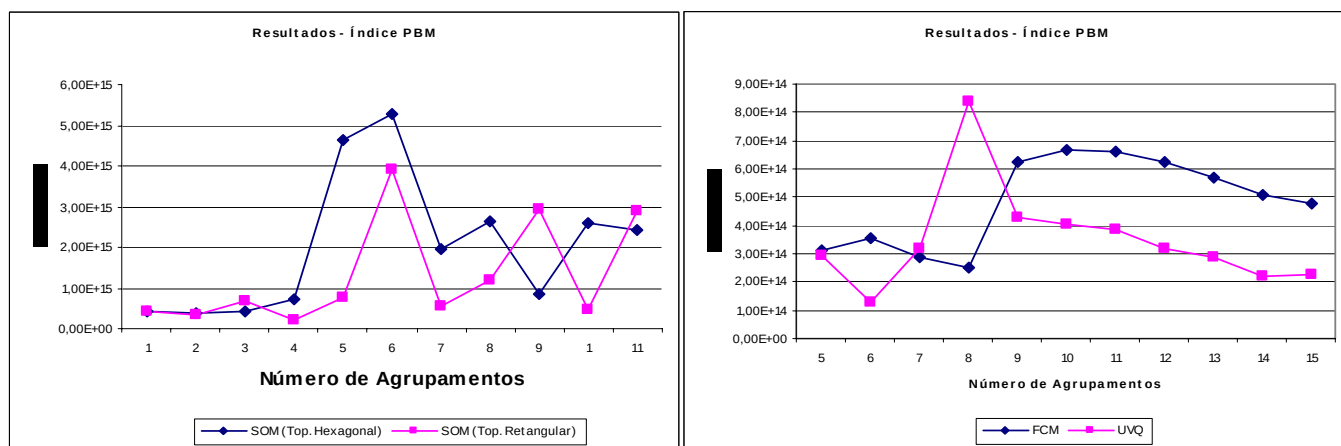


Figura 2 - Resultados da Aplicação do índice PBM

Como pode ser observado pelas tabelas e gráficos apresentados, os melhores resultados obtidos pela aplicação dos índices PBM à rede UVQ apontou para oito agrupamentos (classes), isto é, segundo o índice PBM, o conjunto de dados de aplicação é melhor agrupado em oito classes. Na rede SOM e no algoritmo FCM, os maiores índices apontam para dez classes, tendo a rede SOM apresentado o maior índice geral ($5,29E +15$), utilizando-se a topologia hexagonal.

6. Conclusões

Este trabalho teve seu enfoque na aplicação e comparação de diferentes ferramentas e métodos de agrupamento de dados. Esta foi realizada através da aplicação do índice de validação de agrupamentos PBM, que indicou o melhor agrupamento em cada algoritmo.

Algumas rotinas e funções foram implementadas no programa Matlab enquanto outros programas foram também utilizados. As redes neurais utilizadas correspondem a UVQ e SOM. A rede UVQ faz parte do programa OpendTect. Este programa também foi utilizado para a geração dos atributos sísmicos. A rede SOM faz parte do programa SOM_PAK. O algoritmo FCM foi implementado no programa Matlab, assim como o índice de validação PBM.

Segundo o índice de validação de agrupamentos PBM, os melhores resultados correspondem a oito agrupamentos na rede UVQ e a dez agrupamentos na rede SOM. O resultado da rede UVQ foi exibido em forma de janela sísmica juntamente com os atributos sísmicos utilizados nos conjuntos de dados. A janela sísmica que corresponde ao melhor resultado exibe detalhes não encontrados nas janelas dos atributos sísmicos em separado.

Não se pode afirmar que os agrupamentos representem diferentes litologias. Entretanto, pode-se afirmar que o resultado acima garante, segundo o índice de validação PBM, uma qualidade de agrupamento. A partir deste momento, um especialista poderia analisar estes dados na certeza de que os agrupamentos não foram realizados sem uma validação.

Agradecimentos

Os autores agradecem à ANP por ceder a base de dados utilizada neste trabalho e às agências de fomento CAPES, CNPq e FINEP pelo suporte financeiro ao projeto. Um agradecimento especial à Márcia Khun e aos demais funcionários do Lab2M pelo auxílio nas áreas de Geologia e Geofísica.

Referências

- AZEVEDO, F. M., BRASIL, L.M., OLIVEIRA, R. C. L. Redes neurais com aplicações em controle e sistemas especialistas, *Visual Books*, 2000.
- BEZDEK, J. C. Pattern Recognition with fuzzy objective function algorithms, *Plenum Press*, 1981.
- BRAGA, A. P., LUDEMIR, T. B., CARVALHO, A. C. P. L. F. Redes neurais artificiais, teoria e aplicação, *LTC*, 2000.
- FAUSETT, L. Fundamentals of neural networks, *Prentice Hall*, 1994.
- FLEK, E.M. , PEDREIRA, C. E., SANTOS, R. Agrupamentos de dados sísmicos através do algoritmo de kohonen, *VI Congresso Brasileiro de Redes Neurais*, 2003.
- HAYKIN, S. Redes neurais, princípios e práticas, *Bookman*, 2001.
- KOHONEN, T. Self-organizing maps, *Springer*, 2001.
- LESKI, J. Towards a robust fuzzy clustering, *Fuzzy sets syst.*, v. 137, n.2, p. 215-233, 2003.
- NIKRAVESH, M., AMIZADEH, F., ZADEH, L. *Soft computing and intelligent analysis in oil exploration*, Elsevier, 2003.
- NIKRAVESH, M., HASSIBI, M. Intelligent reservoir characterization, *Journal of petroleum, science and engineering*, v. 29, 2001.
- PAKHIRA, M. K., BANDYOPADHYAY, S., MAULIK, U. Validity index for crisp and fuzzy clusters, *Pattern Recognition*, June 2004.