3
Agrupamento de dados

Agrupamento de dados sísmicos é o processo em que os atributos sísmicos são separados em grupos de forma que seja possível produzir mapas que evidenciem estruturas geológicas (Yin, 2002). Agrupamentos de dados ("*clustering*") podem ser classificados em dois tipos. No primeiro tipo estão os métodos cujos algoritmos utilizam uma função de custo para a formação dos grupos. Os métodos para agrupamentos de dados sísmicos encontrados na literatura estão incluídos neste primeiro tipo assim como o método proposto nesta tese. No segundo tipo enquadram-se outros modelos de algoritmos, dentre os quais, os seqüenciais e os hierárquicos (Theodoridis e Koutroumbas, 1999).

Uma técnica largamente utilizada para agrupamento de dados é a Quantização Vetorial (QV). A idéia básica da QV é representar os grupos através de um número finito de vetores que se denominam de protótipos. Os pontos são alocados em cada um dos agrupamentos de acordo com a técnica de vizinhos mais próximos. Deste modo, a idéia central é formar k grupos de tal forma que as distâncias entre os elementos dos dados de entrada $x = \{x_1, x_2, x_3, ..., x_m\}$ e um dos valores de referência $w = \{w_1, w_2, w_3, ..., w_k\}$ dos grupos seja mínimo. Então, cada elemento do conjunto x é classificado em um grupo da seguinte maneira:

$$w_x = \min_{\substack{j=1,...m \\ i=1,...k}} (d(x_j, w_i)) \qquad (1)$$

$$\text{onde } x \in [1, \text{k}] \text{ é o grupo escolhido}$$

A busca de uma melhor solução para a determinação dos grupos é definida pela localização das referências que é feita através de um processo iterativo que inclui a reavaliação dos valores das referências w e o cálculo de um erro. Busca-se, a cada iteração a minimização do erro que é determinado pela seguinte expressão:

$$e_j = (\sum_{i=1}^{n_j} | x_i - w_j |^p)^{\frac{1}{p}} \quad \text{para p} \geq 1 \qquad (2)$$

Os valores de *w* são determinados por uma função de custo. Uma função que possibilite o cálculo do valor *w* de maneira que o erro na determinação dos grupos seja mínimo é uma função de custo. Alguns itens devem ser observados para a escolha de uma função de custo. Ela deve ser derivável em todos os pontos e deve ser convergente.

Considere a função seguinte:

$$F_{x_j} : \Re \rightarrow \Re$$

$$w \rightarrow \sum_{j=1}^{m} \frac{(x_j - w)^2}{2} \qquad (3)$$

$$F_x = \sum_{j=1}^{m} \frac{(x_j - w)^2}{2} \qquad (4)$$

Cuja primeira derivada é:

$$F_x^{'}(w) = \sum_{j=1}^{m} (w - x_j) \qquad (5)$$

E, cuja segunda derivada é:

$$F_x^{''}(w) = m \qquad (6)$$

Por (6) podemos dizer que *F* é convexa e, portanto, tem um ponto de mínimo local no ponto onde a primeira derivada é zero. O posicionamento de *w* no ponto mínimo de *F* é encontrado para a primeira derivada nula. Por (5) temos:

$$F_x^{'}(w) = \sum_{j=1}^{m} (w - x_j) = 0 \Rightarrow mw - \sum_{j=1}^{m} x_j = 0 \Rightarrow$$

$$w = \frac{1}{m} \sum_{j=1}^{m} x_j$$

Então, o valor para *w* que coincide com a média do grupo proposto é o ponto que, para uma distribuição de dados, implica um erro mínimo e um ponto de mínimo local da função *F*. Com efeito, se os grupos propostos tem

distribuições normais, a média é o ponto central de distribuições simétricas. Os algoritmos que utilizam a função de custo *F* são ditos algoritmos que induzem w à média.

O processo de formação de grupos, como já mencionado, é iterativo. Uma das técnicas utilizadas para esta finalidade é o Gradiente Descendente. Partindo de um valor inicial aleatório, w, através da técnica Gradiente Descendente, é atualizado na direção do valor mínimo da função *F* utilizada (Bishop, 1995). Em outros termos, determina-se o gradiente do erro que decresce com o cálculo da primeira derivada. A atualização iterativa de cada w, onde *t* indica uma iteração, pode ser expressa da seguinte maneira:

$$w^{t+1} = w^t - h\Delta E \quad (7)$$

Onde: ΔE é o gradiente do erro;

h é o termo de amortecimento ou aprendizagem.

O termo de amortecimento pode ser definido de várias maneiras. Se a escolha por h for variável decrescente, normalmente é definido em função do valor da iteração ou múltiplo da iteração. Por exemplo: $h = \frac{1}{t}$. A escolha de h variável decrescente garante a convergência de *w* para o ponto mínimo de *F*. Na pratica, um valor constante para h determina melhores resultados de agrupamentos. No entanto, esta escolha implica a perda da garantia de convergência de *w* para o ponto mínimo.

O gradiente do erro, como já foi mencionado, é avaliado pela variação da primeira derivada de *F*, ou seja:

$$\Delta E = F_x^{'}(w) = \sum_{j=1}^{m} (w - x_j) \quad (8)$$

A substituição de (5) em (7) deve ser precedida do conceito de época. Cada vez que todos os dados de entrada "*x*" são processados é computada uma época do processo (Haykin, 1999). Então o somatório da equação (8) deve ser adequado para cada iteração. Para tal é necessário considerar um grupo de dados *X* em que $x = \{x_1, x_2, x_3, ..., x_m\}$ seja repetido tantas vezes quanto a quantidade de épocas.

$$X \equiv \{\overbrace{x_1, x_2, ..., x_m}^{época1}, \overbrace{x_1, x_2, ..., x_m}^{época2}, ..., \overbrace{x_1, x_2, ..., x_m}^{épocafinal}\} \quad (9)$$

Então, a substituição de (5) em (7) é expressa, em termos de iterações como a seguir:

$$w^{t+1} = w^t - h(w^t - X^t)$$
$$t \geq 0 \text{ indica a iteração} \qquad (10)$$

A generalização de (10) para $i = \{1,...,k\}$ grupos é:

$$w_c^{t+1} = w_c^t + h(X^t - w_c^t) \qquad (11)$$

Onde:

$$w_c = \min_{i=1,...k}^{l=1,...m}(d(x_j, w_i))$$
$$\text{onde } c \in [1,k] \text{ é o grupo escolhido} \qquad (12)$$

As equações (2), (11) e (12) são à base de vários algoritmos que utilizam funções de custo para agrupamento, dentre eles o SOM de Kohonen (Chakraborty et al., 2001; Kangas et al., 1990). Este método é freqüentemente utilizado para agrupamento de dados sísmicos (Essenreiter *et al.*, 2001; Strecker e Uden, 2002; Tarvainen, 1999).

Como foi visto, a média é o estimador para a busca da melhor classificação para algoritmos que utilizam as equações citadas.

Os métodos de agrupamento que tratam distribuições assimétricas são evidentemente menos eficientes se utilizam médias como estimador do centro dos grupos formados (Pitas et al. 1996; Jacobs et al.,2000). Estes métodos, por exemplo, os baseados nos algoritmos de quantização vetorial na sua forma clássica, são não robustos contra ruídos e *"outliers"* (Kamgar-Parsi, et *al.* 1989; Stewart, 1997).

Outro estimador existente para as referências dos grupos propostos é a mediana.

Não foi encontrado, na literatura pesquisada, algoritmo com função de custo que induza as referências dos grupos à mediana.

Na literatura são encontrados algoritmos que atualizam as referências *w* através do computo dos valores da mediana do grupo proposto. Pitas et al. (1996) propõem um modelo VQ com uma atualização de *w* baseada no cálculo da mediana. O processo pode ser sumariado da seguinte maneira:

1. Seja Z_i^t um vetor com os elementos já selecionados do grupo i na iteração t;
2. Seja x um elemento do conjunto de origem X que esta sendo classificado no grupo Z_i^{t+1} na iteração $t+1$;
3. Então, a atualização de w é realizada obedecendo o seguinte critério: $w^{t+1} = mediana\{x \cup Z_i^t\}$.

Kohonen e Somervuo (2002) ajustam a idéia da mediana do vetor sugerida por Pitas et al. (1996) para o SOM. A proposta destes autores é:

1. Seja Z_i^t um vetor com os elementos já selecionados do grupo i na iteração t;
2. Seja x um elemento do conjunto de origem X que esta sendo classificado no grupo Z_i^{t+1} na iteração $t+1$;
3. Seja Y_i^t o conjunto de todos os vizinhos de Z_i^t inclusive o próprio Z_i^t;
4. Então, a atualização de w é realizada obedecendo o seguinte critério: $w^{t+1} = mediana\{x \cup Y_i^t\}$.

Georgakis et al. (2004) elaboraram uma proposta a partir da idéia de Kohonen e Somervuo (2002) propondo diferente forma de atualizar os conjuntos vizinhos ao w vencedor. A lista de atividades que resume a proposta de Georgakis et al. (2004) é:

1. Seja Z_i^t um vetor com os elementos já selecionados do grupo i na iteração t;
2. Seja x um elemento do conjunto de origem X que esta sendo classificado no grupo Z_i^{t+1} na iteração $t+1$;
3. Seja Y_i^t o conjunto de todos os vizinhos de Z_i^t;
4. Então, a atualização de w para o grupo vencedor é realizada obedecendo o seguinte critério: $w^{t+1} = mediana\{x \cup Z_i^t\}$.
5. A atualização de w para os grupos vizinhos é realizada obedecendo o seguinte critério: $w^{t+1} = mediana\{x \cup aY_i^t\}$ onde a é um atenuador que decai com o incremento das épocas.

A idéia de calcular as referências *w* à mediana na forma proposta pelos autores acima garante o valor de *w* à mediana dos grupos formados. Resultados numéricos obtidos nos testes, descritos no capítulo de resultados desta tese, mostram que a classificação de grupos na aplicação deste método aos dados sísmicos simulados tem percentual de erro maior do que os percentuais do algoritmo proposto apresentada nesta tese. A busca da mediana é importante na solução do tratamento do ruído. No entanto, considerando que os dados sísmicos são em grande quantidade, o tempo dos processos acima relacionados é considerável para cada um dos métodos existentes.

Em resumo podemos afirmar:

1. Algoritmos que determinam as referências dos grupos através do cálculo da mediana requerem tempo de processo considerável e, como pode ser visto no capítulo de resultados, os algoritmos existentes não apresentam vantagem na classificação de dados assimétricos vis-à-vis o algoritmo proposto.
2. A literatura que aborda o agrupamento dos dados sísmicos faz referência aos métodos que utilizam algoritmos que induzem as referências dos grupos à média;
3. A média não é um estimador robusto aos ruídos, característica existente nos dados sísmicos;
4. A média não é um estimador para as referências dos grupos propostos em distribuições assimétricas.

A figura 6 mostra as interligações dos assuntos da tese descritas neste capítulo.

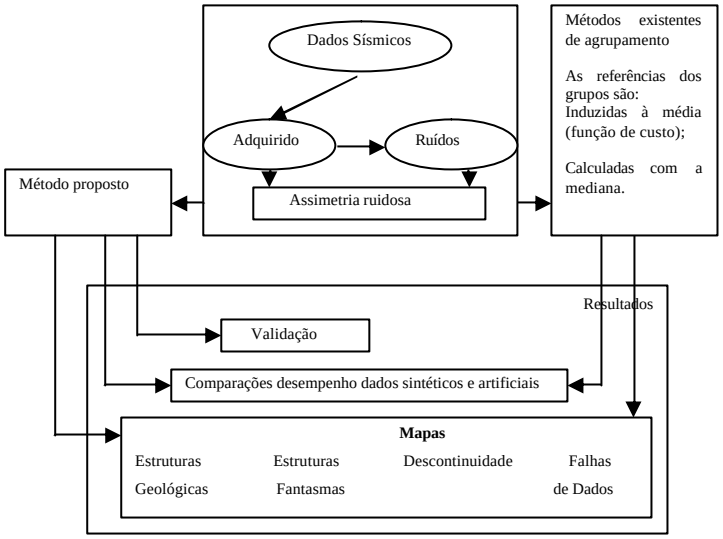

Figura 06 – Interligações dos assuntos da tese – capítulo 3

A opção de indução das referências dos grupos à mediana em método que forneça detalhamento de estruturas geológicas e nitidez na visualização desta estruturas nos mapas é uma lacuna que a proposta desta tese soluciona. O método proposto é descrito no próximo capítulo.