4
Método de indução das referências às medianas

Como foi descrito na introdução, o objetivo central desta tese é a proposta de um novo método para agrupar dados sísmicos para a visualização em mapas. O método utiliza os preceitos de Redes Neurais Artificiais não supervisionadas, particularmente uma variação do método dos vizinhos mais próximos com atualização de protótipos através de algoritmo tipo quantização vetorial (VQ).

O método proposto nesta tese engloba procedimentos para a determinação de grupos em dados sísmicos objetivando a visualização em mapas geológicos. A descrição do método inclui:

- Detalhamento do algoritmo;
- Tarefas pré-processamento do algoritmo;
- Tarefas pós-processamento do algoritmo.

O método proposto consiste da inclusão na equação (11)

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

de um fator b definido como:

$$b \equiv 1 - \frac{c(2s^2 - (X - w)^2)}{s^4} \exp\left(\frac{-(X - w)^2}{2s^2}\right) \qquad (13)$$

A equação resultante é:

$$w^{t+1}(x) = w^t(x) + hb^t(X^t - w^t) \quad (14)$$

que acarreta deslocamentos no posicionamento das referências para a média ou para a mediana ou, ainda, para valores próximos destes dois estimadores. O método possibilita o agrupamento de dados sísmicos robusto aos ruídos de maneira a proporcionar uma visualização de estruturas geológicas detalhadas e nítidas.

4.1
Análise de b

O fator b engloba três termos: a distância entre os elementos e as referências dos grupos *(X – w)*, o desvio padrão s e uma constante de não linearidade *c* válida para todo o processo.

Para processos com mais de uma variável, a determinação das distâncias requer atenção especial. Nesta tese a distância estatística (Johnson e Wichern, 1998) é a escolhida. Esta distância é determinada através do cálculo da distância Euclidiana em que os termos são divididos pelo desvio padrão de cada variável envolvida no processo. Por exemplo, a distância estatística entre a origem e um ponto P definido por duas variáveis $P = (x_1, x_2)$ é:

$$d(O,P) = \sqrt{(\frac{x_1}{\sqrt{s_{11}}})^2 + (\frac{x_2}{\sqrt{s_{22}}})^2} = \sqrt{\frac{x_1^2}{s_{11}} + \frac{x_2^2}{s_{22}}} \quad (15)$$

Os valores s_{11} e s_{22} representam os desvios padrões de cada variável.

Outro termo do fator b é o desvio padrão. Para processos multivariáveis, o desvio padrão é representado pela raiz quadrada da matriz de variância – covariância S_n de uma população de n observações. Esta matriz é calculada, para p variáveis onde m_i representa a média da variável $i \in [1, p]$ da seguinte forma:

$$S_n = \begin{bmatrix} \frac{1}{n}\sum_{j=1}^{n}(x_{j1} - m_1)^2 & \cdots & \frac{1}{n}\sum_{j=1}^{n}(x_{j1} - m_1)(x_{jp} - m_p) \\ . & & \\ . & & \\ . & & \\ \frac{1}{n}\sum_{j=1}^{n}(x_{j1} - m_1)(x_{jp} - m_p) & \cdots & \frac{1}{n}\sum_{j=1}^{n}(x_{jp} - m_p)^2 \end{bmatrix} \quad (16)$$

Na diagonal da matriz S_n são calculadas as variâncias de cada variável. Os outros elementos desta matriz fazem referências às covariâncias. Covariâncias nulas refletem variáveis independentes. Este é o cenário ideal para processos de agrupamento. No entanto, o método permite a utilização de variáveis com diferentes níveis de covariâncias acarretando, no caso destas covariâncias serem altas, poucas alterações nos resultados do processo de classificação. Para

processos univariáveis, caso desenvolvido nesta tese, o desvio padrão é referenciado como s .

A constante de não linearidade c determina a magnitude do seguinte termo à direita de b :

$$\frac{c(2s^2-(X-w)^2)}{s^4} \quad .$$

Este termo, dependendo do valor de c, é submetido a um filtro Gaussiano que é representado pelo segundo termo da direita de b . O fator b pode ser visto, então, como o complemento da combinação dos dois termos.

Fazendo $d=(X-w)$ e multiplicando a expressão acima por $\frac{2}{s^2}*\frac{s^2}{2}$ temos:

$$\frac{2c}{s^2}\left(\frac{2s^2}{s^4}*\frac{s^2}{2}-\frac{d^2}{s^4}*\frac{s^2}{2}\right)=\frac{2c}{s^2}\left(1-\frac{d^2}{2s^2}\right)$$

Fazendo $x=\frac{d}{\sqrt{2}s}$ é possível reescrever b em função de x da seguinte forma:

$$b(x)=1-\frac{2c}{s^2}(1-x^2)e^{-x^2} \qquad (17)$$

A representação gráfica do fator b para uma mistura de distribuições normais é mostrada na figura 6 (lado esquerdo).

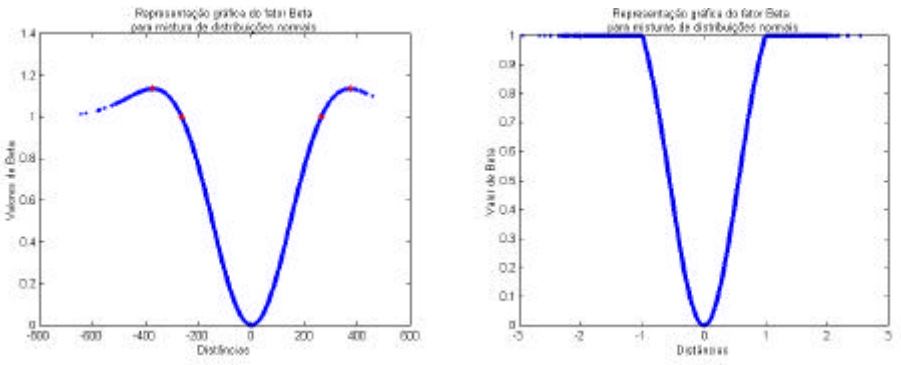

Figura 6 – Representação gráfica do fator b para misturas de distribuições normais

A representação gráfica deste fator é semelhante ao "chapéu mexicano" invertido. Se considerarmos a unidade como valor máximo utilizado para b, a figura 6 (lado direito) mostra como os valores de b atualizam a equação (14).

A primeira derivada de b é:

$$b'(x) = \frac{2c}{s^2}(2 - x^2)2xe^{-x^2} \quad (18)$$

Os pontos de mínimo e máximo de b são os correspondentes aos valores de x que fazem $b' = 0 \Rightarrow$ $x = 0$ (ponto de mínimo) $x = \pm\sqrt{2}$ (pontos de máximos)

E, cuja segunda derivada é:

$$b'' = \frac{2c}{s^2}(4 - 14x^2 + 4x^4)e^{-x^2} \quad (19)$$

A figura 7 mostra as representações gráficas das duas primeiras derivadas de uma mistura de distribuições normais. Considerando-se os limites $x = \pm\sqrt{2}$ (pontos máximos de b) temos que o ponto zero é um ponto de mínimo absoluto pois, na primeira derivada:

$$b'(x) < 0 \qquad \forall x < 0$$
$$b'(x) > 0 \qquad \forall x > 0$$

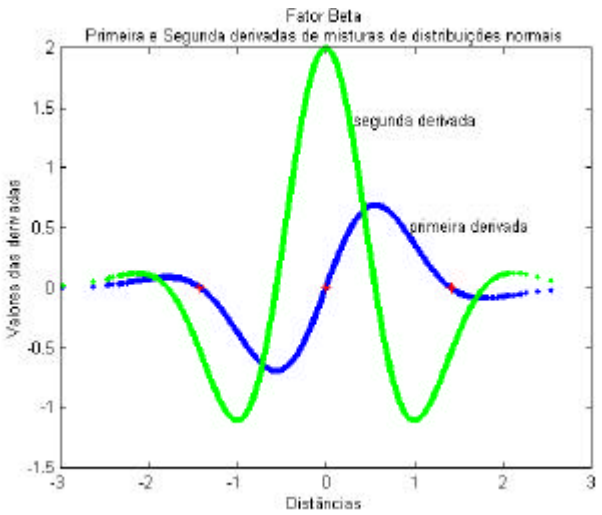

Figura 7 – Representação gráfica das derivadas do fator b para misturas de distribuições normais

As características de b são enunciadas abaixo como proposições.

Proposição 1

b é uma função simétrica em x.

Prova:

Pela equação (17) b é uma função simétrica em x.

Proposição 2

O valor de X para o qual b é mínimo é igual a referência do grupo. Em outras palavras, o valor mínimo do fator b é sempre coincidente com o valor da referência do grupo.

Prova:

O ponto de valor mínimo de b é fornecido pela equação (18)

$b' = 0 \Rightarrow \quad x = 0$ (ponto de mínimo)

Como $x = \frac{d}{\sqrt{2}s}$ o valor mínimo de b corresponde a distância $d = 0$. Em outros termos, $b_{\min} \Rightarrow d = 0 = (X - w)$. Então, $b_{\min} \Rightarrow X = w$.

Proposição 3

O valor mínimo de b é inversa e linearmente relacionado com *c*.

Prova:

O valor de b para o ponto de mínimo é dado por:

$$b(0) = 1 - \frac{2c}{s^2}(1-0)e^0$$

$$b(0) = 1 - \frac{2c}{s^2}$$

$$b_{\min} = 1 - \frac{2c}{s^2} \quad (20)$$

Proposição 4

O valor máximo da constante c, correspondente ao valor mínimo de $b = 0$, é $c = \frac{s^2}{2}$.

Prova:

Os valores de b não podem ser negativos. Com efeito, valores negativos para b na equação (14) certamente farão com que a direção do gradiente seja contrária ao movimento de convergência à um estimador escolhido, seja este a média ou a mediana. O valor $b = 0$ é, portanto, a opção para o menor valor mínimo de b.

Então, pela equação (20) temos:

$$b_{\min} = 1 - \frac{2c}{s^2}$$

$$0 = 1 - \frac{2c}{s^2}$$

$$c = \frac{s^2}{2} \quad (21)$$

Proposição 5

O valor máximo de $b = 1 + \frac{0.2706c}{s^2}$.

Prova:

Os valores de b para os pontos de máximas são:

$$b(\pm\sqrt{2}) = 1 - \frac{2c}{s^2}(1-2)e^{-2}$$

$$b_{\max} = 1 + \frac{0.2706c}{s^2}$$

$$b(\pm\infty) = 1$$

A avaliação da equação (14) permite verificar que os valores de $b > 1$ (por exemplo, $b_{\max}$) podem induzir divergências dos valores das referências dos grupos à qualquer estimador escolhido seja ele a média ou a mediana. A solução de contorno para este caso é fixar o valor máximo de $b_{\max} = 1$. Com efeito, o valor de $b_{\max} = 1$ não só contempla uma condição necessária para a convergência do processo como permite a utilização do método proposto para a convergência das referências ao estimador escolhido. Para a média, por exemplo, basta fixar o valor de $b = 1$ para todo o processo.

Proposição 6

Se $b = 1$ então $-\sqrt{2}s \leq (X - w) \leq +\sqrt{2}s$.

Prova:

Pela equação (17) $b = 1 - \frac{2c}{s^2}(1 - x^2)e^{-x^2}$

$$1 = 1 - \frac{2c}{s^2}(1 - x^2)e^{-x^2} \Rightarrow x^2 = 1.$$

Mas,

$$x = \frac{d}{\sqrt{2}s} = \pm 1 \Rightarrow d = \pm\sqrt{2}s = (X - w).$$

4.2. Indução à mediana

As proposições listadas no item anterior servem para melhor compreender a determinação das constantes de atenuação h e de não linearidade c. As experiências empíricas com diversos tipos de distribuições de dados assimétricas (misturas de distribuições normais, distribuição exponencial, distribuição aleatória e distribuição aleatória enxertada com *"outliers"*) mostram a convergência à mediana da referência das distribuições examinadas.

4.3. Tarefas pré-processamento do algoritmo

No caso da distribuição dos dados de entrada ser normal:

- Os dados sísmicos com distribuições normais devem ter $c = 0$. É recomendável que o valor de h para estes casos seja variável decrescente.

No caso da distribuição dos dados de entrada ser assimétrica:

- O cálculo do desvio padrão dos dados de entrada é a primeira tarefa a ser realizada;
- Determinação de c e h. O valor de c determina o quanto à função de custo é não linear. Para uma mesma constante h, valores de c que determinam baixa não linearidade fornecem desvios padrões

menores para os grupos propostos. O aumento na direção de $c = \frac{s^2}{2}$ acarreta um aumento de desvio padrão nos grupo. Os mapas com c no entorno de $\frac{s^2}{4}$ apresentam figuras mais continuas definindo estruturas geológicas bem delineadas. Na direção de uma menor não linearidade, os mapas são mais detalhados apresentando figuras com maior descontinuidade nas estruturas geológicas. O valor $c = \frac{s^2}{4}$ é recomendado como valor inicial para os processos com dados assimétricos. A escolha de um valor constante de $0 < h < 1$ é recomendado. A escolha de $0 < h < 1$ constante implicou menores percentuais de erro nos testes numéricos elaborados e mostrados no capítulo de resultados.

4.4. Tarefas pós-processamento do algoritmo

A tarefa pós-processamento que antecede a visualização dos mapas é a avaliação do erro do processo. O algoritmo converge rapidamente para um bom resultado. No entanto, é importante avaliar o descaimento do erro durante o processo para determinação, se necessário, de nova parametrização dos valores de c e h ou uma quantidade maior de épocas para a determinação de um erro menor.