4
Casamento de Padrões

O casamento de padrões é uma técnica que tem por objetivo localizar os elementos constituintes de uma seqüência em um conjunto de outras seqüências. Chamemos de padrão a seqüência de elementos que se deseja buscar e de alvo uma seqüência de elementos no conjunto de seqüências formadoras da base de busca.

Quando os elementos constituintes das seqüências são símbolos discretos, como caracteres, e se deseja exatamente a posição onde o padrão ocorre numa seqüência alvo, algoritmos como Knuth-Morris-Pratt (Cormen et al., 2002) ou Boyer-Moore (Boyer e Moore, 1977) podem ser usados.

A abordagem muda quando as seqüências são formadas por valores de ponto flutuante. Pode-se tentar usar um dos algoritmos mencionados, o que requer uma discretização dos valores dos elementos constituintes ou a determinação de limiares para a verificação de igualdade entre elementos. Estas restrições adicionais tornariam o problema de casamento de seqüencias de ponto flutuante mais complexo, uma vez que, para cada domínio de utilização destes algoritmos seria necessário fornecer as restrições.

Em algumas ocasiões o padrão pode estar contido em um alvo de forma não contínua, ou seja, pode haver saltos entre as posições de casamento entre padrão e o alvo. Além disso, o casamento pode não ser completo e não ser exato. Por casamento completo entenda-se o casamento em que todos os elementos constituintes do padrão se encontram, em ordem, na seqüência alvo. Casamento não exato pode ser entendido como o pareamento entre padrão e alvo de tal forma que é admitida uma diferença entre elementos casados. Para o caso de casamento incompleto, mas exato, o algoritmo *longest common subsequence* ou, em tradução livre, *maior subseqüência em comum* pode ser utilizado (Cormen et al., 2002).

O caso de estudo deste trabalho é o casamento entre uma curva, o traço sísmico sintético, ao longo do caminho de um poço, e a sísmica 3D. Ambos, sísmica 3D e poço ou poços devem pertencer ao mesmo sistema de coordenadas para que seja possível realizar o casamento. Valores na sísmica e no traço sintético são, freqüentemente, ponto flutuante, e uma abordagem não exata se torna necessária.

4.1
Casamento de variância mínima

Em Latecki et al. (2007) é mostrado um algoritmo de casamento parcial de formas chamado de *minimum variance matching* (MVM) ou, em tradução livre, casamento de variância mínima. Este algoritmo mapeia o problema de subseqüência de melhor casamento ao problema de caminho mais barato em um grafo direcionado acíclico (DAG), transformando as formas ou curvas em seqüencias numéricas para as quais o algoritmo MVM seja aplicável.

Seja $p = (p_1, \ldots, p_m)$ o padrão de busca e $t = (t_1, \ldots, t_n)$ a seqüência alvo, com $m < n$. Uma correspondência entre p e t pode ser entendida como como uma injeção monotônica crescente $f : \{1, \ldots, m\} \rightarrow \{1, \ldots, n\}$, ou seja, um mapeamento de p_i para $t_{f(i)}$, para todo $i \in \{1, \ldots, m\}$. O objetivo do algoritmo MVM é encontrar a melhor correspondência $\hat{f}$ dentre as correspondências f tal que a dissimilaridade entre os elementos casados seja mínima. Para medir a dissimilaridade, pode ser usada, por exemplo, a distância euclidiana. Assim, a dissimilaridade entre o padrão e o alvo, dada a correspondência, pode ser dada por:

$$d(p, t, f) = \sqrt{\sum_{i=1}^{m} (t_{f(i)} - p_i)^2}. \tag{4-1}$$

Daí, a melhor correspondência é

$$\hat{f} = \arg\min\{d(p, t, f) : f \text{ é uma correspondência}\},$$

e a dissimilaridade da melhor correspondência, ou o mínimo global dentre todas correspondências é

$$d(p, t) = d(p, t, \hat{f}) = \sqrt{\sum_{i=1}^{m} (t_{\hat{f}(i)} - p_i)^2}. \tag{4-2}$$

O algoritmo MVM é baseado em programação dinâmica (Kleinberg e Tardos, 2005), e sua execução se inicia pela geração de uma matriz de diferenças entre pares de elementos do padrão e do alvo, como mostra a Figura 4.1.

A partir da matriz de diferenças é criado um grafo direcionado acíclico. Cada elemento da matriz de diferença r é encarado como um nó no grafo, e há uma aresta do nó r_{ij} ao nó r_{kl} se $k - i = 1$ e $j + 1 \leq l \leq j + n - m$. O custo de cada aresta é dado por:

$$custo(r_{ij}, r_{kl}) = \begin{cases} (r_{kl})^2 = (t_k - p_i)^2 & \text{se } k = i + 1 \text{ e } j + 1 \leq l \leq j + n - m, \\ \infty & \text{em outro caso} \end{cases}$$

$$r = \begin{bmatrix} (0) & 1 & 8 & 2 & 2 & 4 & 8 \\ 1 & (0) & 7 & 1 & 1 & 3 & 7 \\ -7 & -6 & (1) & -5 & -5 & -3 & 1 \\ -5 & -4 & 3 & -3 & -3 & (-1) & 3 \\ -7 & -6 & 1 & -5 & -5 & -3 & (1) \end{bmatrix}$$

Figura 4.1: Matriz de diferenças entre o padrão $p = (1, 2, 8, 6, 8)$ e o alvo $t = (1, 2, 9, 3, 3, 5, 9)$ formada com linhas correspondentes a elementos de p e colunas a elementos de t ($r_{ij} = t_j - p_i$). Em parênteses, a melhor correspondência segundo o algoritmo MVM. Adaptada de (Latecki et al., 2007).

Adicionalmente, é necessário criar um nó raiz com arestas para os nós r_{1j}, com $j \in \{1, \ldots, n\}$ cujo custo é $(r_{1j})^2$. Criado o grafo direcionado acíclico, pode-se aplicar o algoritmo de caminho mais barato em um DAG para retornar todos os caminhos com seus respectivos custos a partir do nó raiz (Cormen et al., 2002). Dentre a lista de caminhos, pode-se, então, selecionar aquele que tem o menor custo total associado.

O algoritmo MVM serve para casamento parcial elástico, ou seja, são permitidos intervalos inteiros sem casamento entre o alvo e o padrão, significando que pode ser necessário "esticar" o padrão a fim de que haja o melhor casamento possível. A este tamanho do intervalo sem casamento os autores do algoritmo MVM chamam de elasticidade, uma propriedade global aplicada ao padrão para que esta case com o alvo.

As Tabela 4.1 mostra o efeito de diferentes valores de elasticidade a partir de uma seqüência e de um padrão.

Seqüência:	2	3	1	5	8	9	3	2	8	0	1	2	4	6	7	9	1	2	7	1
$e = 0$		4	1	5	9	8	2	7	1											
$e = 1$		4	1	5	9	8		2	7		1									
$e = 2$		4	1	5		9			8			2			7		1			

Tabela 4.1: Efeito de diferentes elasticidades sobre o casamento entre uma seqüência e o padrão $\{4, 1, 5, 9, 8, 2, 7, 1\}$. Cada emparelhamento mostrado abaixo da seqüência é resultado do algoritmo MVM usando diferentes valores de elasticidade (e).

A Figura 4.2 mostra um exemplo de casamento entre duas curvas com valores em ponto flutuante, tendo a curva alvo 2500 amostras e a curva de busca (padrão) 500 amostras. A elasticidade global utilizada tem valor 2.

4.2

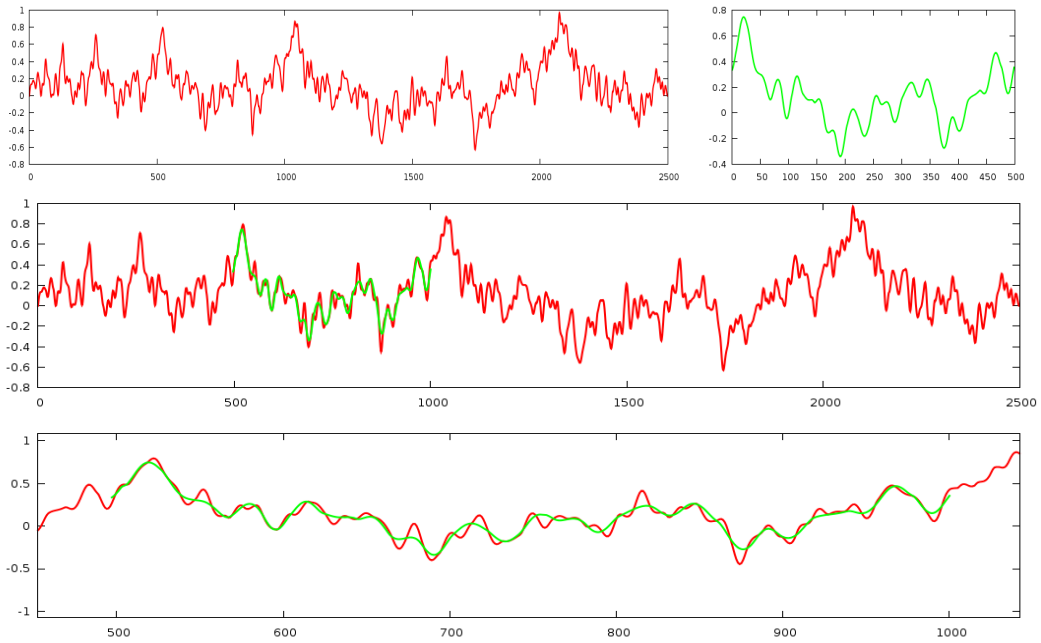

Figura 4.2: Casamento de curvas usando o algoritmo MVM. Acima à esquerda, a curva alvo. Acima à direita, curva de busca (padrão). Ao centro, casamento realizado. Abaixo, detalhe do casamento. Curvas geradas usando Perlin Noise (Perlin, 1985).

Casamento de variância mínima com múltiplas elasticidades

Nesta dissertação, a fim de contemplar múltiplos intervalos de coincidência parcial elástica, é proposta uma mudança no algoritmo MVM original de forma que seja aceita uma lista de valores de elasticidade, de forma a manter grupos de amostras mais fortemente ligados, com valores de elasticidade pequenos, e a separar grupos de amostras independentes, com valores de elasticidade grandes (maiores que o tamanho do padrão). A mudança é implementada atribuindo a cada posição do padrão um valor de elasticidade, que é o tamanho máximo do intervalo sem casamento até a próxima amostra.

Os valores de elasticidade aplicados a cada trecho devem ser fornecidos *a priori*, servindo como um controle do casamento esperado entre o padrão a seqüência.

Uma vantagem do uso de múltiplas elasticidades é a de resolver um problema combinatorial de posicionamento de trechos de padrão na seqüência mantendo a complexidade original do algoritmo. Executando o algoritmo MVM original usando como padrão cada trecho a ser buscado, poderíamos chegar a uma situação em que os trechos casados com a seqüência poderiam ou não estar na ordem ou na posição desejadas, como mostra o exemplo da Tabela 4.2.

Parte anterior do padrão (A):						4	1	5	9											
Parte posterior do padrão (B):						3	8	3	8											
Seqüência:	2	3	1	5	8	2	3	2	8	0	1	2	4	1	7	9	1	2	7	1
$e = 0$ (A):		4	1	5	9															
$e = 1$ (A):		4	1	5	9															
$e = 2$ (A):		4	1	5	9															
$e = 3$ (A):		4	1	5	9															
$e = 0$ (B):												3	8	3	8					
$e = 1$ (B):			3		8		3		8											
$e = 2$ (B):		3			8		3		8											
$e = 3$ (B):		3			8		3		8											

Tabela 4.2: Resultado de execuções do algoritmo MVM para os trechos (A) e (B) em separado usando variados valores de elasticidade. Percebe-se sobreposição entre os casamentos dos trechos (A) e (B) para elasticidade igual a 1, 2 e 3.

Para eliminar o problema mostrado na Tabela 4.2, seria necessário definir regiões dentro das quais os padrões deveriam ser casados com a seqüência, o que também elimina parte das combinações possíveis de posicionamento de trechos de padrão em relação às posições da seqüência. No exemplo da Tabela 4.3, são criados dois grupos de amostras, e a separação entre estes dois grupos ocorre quando da presença de uma elasticidade grande. Desta forma, a complexidade do algoritmo MVM é mantida e, ao mesmo tempo, consegue-se resolver o problema combinatorial de posicionamento dos grupos de amostras do padrão na seqüência de busca.

Padrão:				4	1	5	9	8	2	7	1									
Tabela de elasticidades:				1	1	1	∞	1	1	1	-									
Seqüência:	2	3	1	5	8	9	3	2	8	0	1	2	4	6	7	9	1	2	7	1
Casamento:		4	1	5		9										8		2	7	1

Tabela 4.3: Efeito da modificação aplicada ao algoritmo MVM original. O padrão é dividido em dois grupos ({4,1,5,9} e {8,2,7,1}) pela presença de um número muito grande na posição final do primeiro grupo de amostras.

4.3 Análise de complexidade

Seja m o tamanho do padrão e n o tamanho da seqüência de busca. A criação da tabela de diferenças r tem complexidade $O(nm)$, mas somente há a necessidade de criar no máximo $m(n-m)$ nós no grafo. Todo vértice r_{ij} na linha i está ligado a no máximo $n-m-j+1$ vértices na linha $i+1$. Como

$$\sum_{j=1}^{n-m+1} (n-m-j+1) = \frac{(n-m)(n-m+1)}{2}$$

e existem m linhas, então o DAG tem no máximo

$$m \times \frac{(n-m)(n-m+1)}{2}$$

arestas. Como encontrar o caminho mais curto num DAG tem custo $O(V+E)$, sendo V o número de vértices e E o número de arestas, então o MVM tem complexidade $O(mn^2)$. Entretanto, esta complexidade pode ser reduzida para $O(mn)$ se for considerado o fator de elasticidade (Latecki et al., 2007).

A modificação aplicada ao algoritmo MVM original não aumenta sua complexidade original, uma vez que a lista de elasticidades pode ter, no pior caso, valores muito grandes, igualando sua complexida à mesma da análise inicial do algoritmo MVM sem considerar a elasticidade. Assim, a complexidade do algoritmo MVM modificado é também $O(mn^2)$.