ESTUDO DO COMPORTAMENTO DO ALGORITMO DE FORÇA BRUTA NA IDENTIFICAÇÃO DE MOTIFS

Jhonny Marcos Acordi Mertz (outros/Unioeste), Huei Diana Lee, Wu Feng Chung, André Gustavo Maletzke (Orientador), e-mail: andregustavom@gmail.com

Universidade Estadual do Oeste do Paraná/ Centro de Engenharias e Ciências Exatas/ Campus de Foz do Iguaçu/ PR

Palavras-chave: séries temporais, padrões, mineração de dados.

Ciências Exatas e da Terra - Ciência da Computação

Resumo

O crescente interesse pela extração de conhecimento a partir de dados que possuam características temporais tem possibilitado a solução de distintos problemas, os quais não poderiam ser solucionados sem que fosse considerado o fator tempo. Nesse sentido, a busca por padrões em séries temporais pode auxiliar na análise desses dados, pois esses padrões podem descrever a série. Neste trabalho, é realizado a avaliação e estudo da abordagem de força bruta para identificação de padrões em séries temporais. Os estudos indicam que a abordagem é eficaz na identificação de padrões, no entanto, requer um alto custo computacional para ser aplicada.

Introdução

Atualmente, cada vez mais, distintos domínios têm demonstrado interesse na monitoração de fenômenos ao longo do tempo. Dados coletados por meio de processos de monitoração apresentam características temporais e, comumente, são representados por meio de Séries Temporais (ST) [1, 5].

Entretanto, os métodos tradicionais de análise de dados possuem restrições a dados que apresentam características temporais, devido à informação temporal não ser considerada no processo de análise. Nesse sentido, torna-se necessário estudar e desenvolver abordagens alternativas para análise de dados temporais. Desse modo, uma área de interesse dentro da mineração de ST se refere à busca de padrões morfológicos frequentes. Esses padrões, denominados *motifs,* são previamente desconhecidos e se repetem ao longo de uma ST podendo constituir atributos importantes na descrição desses dados [2].

Este trabalho tem como objetivo apresentar uma avaliação do método de identificação de *motifs* denominado Força Bruta (FB) e é parte do projeto Análise Inteligente de Dados, desenvolvido entre o Laboratório de

Bioinformática (LABI) da Universidade Estadual do Oeste do Paraná (UNIOESTE)/Foz do Iguaçu, o Laboratório de Inteligência Computacional (LABIC) da Universidade de São Paulo (USP)/São Carlos, o Grupo Interdisciplinar em Mineração de Dados e Aplicações (GIMDA) da Universidade Federal do ABC (UFABC)/Santo André e o Serviço de Coloproctologia da Faculdade de Ciências Médicas da Universidade Estadual de Campinas (UNICAMP)/Campinas.

Materiais e Métodos

A identificação de *motifs* por meio da abordagem de FB requer grande esforço computacional, pois possui complexidade quadrática $O(n^2)$ em relação ao tamanho da série, considerando uma série temporal como um conjunto de n valores ordenados ao longo do tempo [2, 4]. No entanto, essa abordagem é amplamente utilizada, pois apresenta resultados satisfatórios para ST com baixo número de observações [1].

A identificação de *motifs* consiste em estudar pequenas porções de uma ST, denominadas subsequências. Nesse sentido, o método de FB realiza a comparação de cada subsequência, de tamanho m, de uma ST de tamanho n, para $m << n$, com as subsequências restantes da ST. Para cada comparação verifica-se a ocorrência de casamento, isto é, verifica-se se as subsequências são similares. Para determinar a similaridade entre duas subsequências foram utilizados a distância Euclidiana e um limiar de aceitação igual à zero. Caso sejam similares, suas localizações são armazenadas, pois indicam a existência de um padrão morfológico.

No entanto, antes de verificar a ocorrência de casamento entre duas subsequências é necessário que ambas sejam normalizadas em relação ao eixo das ordenadas, modificando assim, os valores originais, porém preservando a morfologia das subsequências. Para essa tarefa realizou-se a subtração de cada valor da subsequência pelo valor da primeira observação.

Todavia, o processo está suscetível à ocorrência de casamentos triviais, os quais são indesejáveis, pois se referem à ocorrência casamentos entre subsequências vizinhas, com diferença de poucas observações à direita ou esquerda [4]. O algoritmo utilizado contorna esta situação por meio da busca pelo melhor casamento entre subsequências vizinhas.

Com o intuito de avaliar a eficácia do método de FB e a influência da escolha do tamanho do *motif* foram selecionadas dez ST de distintos domínios, provenientes da base de dados *UCR Time Series Classification/Clustering* [3], conforme descritas na Tabela 1.

Tabela 1 – Resumo das características das ST utilizadas.

Série Temporal	Número de Observações	Série Temporal	Número de Observações
Tide	*8.746*	*Dow Jones*	*26.443*

Widing	*2.500*	*Spot Exrates*	*2.567*
Random Walk	*65.536*	*Astrophysical*	*2.204*
EEG	*7.200*	*Network*	*18.000*
Wind	*6.574*	*S&P500*	*17.610*

Após, para cada ST da Tabela 1 foram aplicados os passos a seguir:

- **Seleção das observações da ST:** foram extraídas as 1.000 primeiras observações, representando a ST experimental *T*;
- **Seleção de *motifs* artificiais:** foram extraídas subsequências de distintos tamanhos, iniciando em 25 até 250 observações com incrementos de 25 observações, de posições aleatórias superiores às 1.000 observações reservadas para representar a ST;
- **Construção do conjunto de ST experimentais:** a ST *T* foi replicada dez vezes de modo que a ST T_1 receba duas ocorrências do *motif* de tamanho 25, a ST T_2 receba duas ocorrências do *motif* de tamanho 50, até a ST T_{10}. As posições de inserção foram estabelecidas aleatoriamente, respeitando posições sobrepostas e o comportamento da ST.

Para avaliação do FB foi considerado o sucesso na identificação dos padrões e o tempo de execução para cada variação do tamanho de *motif*. Os experimentos foram realizados em um ambiente controlado com as configurações: CPU Intel® Core™ 2 Duo 2.20 GHz, 4 GB de RAM e sistema operacional Microsoft ® Windows XP *Service Pack* 3. Para a implementação dos experimentos foi utilizada a linguagem e ambiente estatístico R 2.13.1.

Resultados e Discussão

A partir da análise dos resultados da aplicação do FB observou-se uma precisão de 100% na identificação dos *motifs* inseridos.

Tabela 2 – Resultados da execução do algoritmo de FB.

Série Temporal	Tamanho da subsequência artificial									
	25	**50**	**75**	**100**	**125**	**150**	**175**	**200**	**225**	**250**
Astrophysical	12,38 (0,04)	11,94 (0,03)	11,40 (0,03)	10,60 (0,02)	9,78 (0,02)	8,91 (0,02)	8,01 (0,01)	7,11 (0,02)	6,19 (0,01)	5,34 (0,01)
Dow Jones	12,38 (0,04)	11,91 (0,03)	11,40 (0,02)	10,61 (0,02)	9,79 (0,02)	8,91 (0,02)	8,01 (0,01)	7,12 (0,01)	6,22 (0,01)	5,35 (0,01)
EEG	12,40 (0,04)	11,95 (0,03)	11,41 (0,03)	10,61 (0,02)	9,78 (0,02)	8,91 (0,02)	8,01 (0,02)	7,12 (0,01)	6,22 (0,01)	5,35 (0,01)
Network	12,41 (0,03)	11,94 (0,03)	11,41 (0,03)	10,61 (0,02)	9,79 (0,02)	8,91 (0,02)	8,01 (0,02)	7,12 (0,01)	6,22 (0,01)	5,35 (0,01)
Random Walk	12,41 (0,04)	11,92 (0,03)	11,32 (0,02)	10,50 (0,02)	9,69 (0,02)	8,83 (0,02)	7,92 (0,01)	6,99 (0,02)	6,11 (0,02)	5,27 (0,01)
Spot Exrates	12,42 (0,03)	11,95 (0,03)	11,33 (0,13)	10,49 (0,02)	9,69 (0,02)	8,82 (0,02)	7,91 (0,02)	6,99 (0,02)	6,14 (0,01)	5,27 (0,01)
S&P500	12,43 (0,03)	11,95 (0,03)	11,33 (0,09)	10,52 (0,02)	9,71 (0,02)	8,84 (0,02)	7,91 (0,01)	7,01 (0,02)	6,14 (0,01)	5,27 (0,01)
Tide	12,42 (0,03)	11,95 (0,03)	11,33 (0,02)	10,52 (0,02)	9,71 (0,02)	8,83 (0,02)	7,92 (0,03)	6,98 (0,01)	6,14 (0,01)	5,27 (0,01)

Wind	12,42 (0,03)	11,94 (0,03)	11,31 (0,03)	10,49 (0,02)	9,70 (0,02)	8,80 (0,02)	7,90 (0,01)	6,99 (0,02)	6,13 (0,01)	5,27 (0,01)
Winding	12,43 (0,03)	11,95 (0,03)	11,31 (0,02)	10,50 (0,02)	9,70 (0,02)	8,83 (0,02)	7,92 (0,01)	7,01 (0,02)	6,14 (0,01)	5,27 (0,01)

Na Tabela 2 são apresentados os tempos médios, em segundos, com seus respectivos desvios padrão das 100 execuções do algoritmo para a identificação de *motifs*, considerando cada uma das ST em relação à variação do tamanho de *motif*. A análise dos resultados permitiu observar a influência do tamanho do *motif* no tempo de execução do FB, o qual é inversamente proporcional ao tamanho de padrão procurado.

Em [1] é apresentada uma avaliação da abordagem de FB, no entanto, é avaliada somente a influência do tamanho da ST na execução de FB. Desse modo, os resultados desta avaliação possibilitaram avaliar o FB de maneira mais completa em função do seu principal parâmetro, o tamanho do *motif*. Por fim, a precisão obtida por meio do FB faz com que este método seja amplamente utilizado, principalmente para fins de comparação com outros métodos.

Conclusões

Esse estudo permitiu constatar a eficácia do FB para identificação de *motifs* em ST. Além de contribuir com um estudo mais completo em relação ao esforço computacional da abordagem. Como trabalho futuro cita-se a aplicação da extração de *motifs* a dados reais, especificamente, da área médica.

Referências

1. CESTARI, D. M.; MALETZKE, A. G.; BATISTA, G. E. A. P. A. Avaliação do algoritmo de força-bruta para a identificação de padrões frequentes em séries temporais. In **Anais do III Congresso da Academia Trinacional de Ciências**, Foz do Iguaçu, 2008, Vol. 1, 1.

2. MALETZKE, A. G. **Uma metodologia para a extração de conhecimento em séries temporais por meio da identificação de *motifs* e da extração de características**. 2009. Dissertação de Mestrado – Instituto de Ciências Matemáticas e Computação, Universidade de São Paulo, São Carlos, 2009.

3. KEOGH, E.; XI, X.; WEI, L.; RATANAMAHATANA, C. A. **The UCR Time Series Classification/Clustering**. Disponível em: <www.cs.ucr.edu/~eamonn/time_series_data/>. Acesso em: 25 jul. 2011.

4. CHIU, B.; KEOGH, E.; LONARDI, S. Probabilistic discovery of time series motifs. In **Proceedings of the 9th International Conference on Knowledge Discovery and Data Mining**, USA, p.493-498, 2003.

5. MORETTIN, P.A; TOLOI, C. M. C. **Análise de Séries Temporais**. 2 ed. São Paulo: Edgard Blücher, 2004.