Proposta de Otimização do Algoritmo de Força Bruta para a Identificação de *Motifs*

Jhonny Marcos Acordi Mertz[1], André Gustavo Maletzke[1], Huei Diana Lee[1,2], Feng Chung Wu[1,2]

[1]Laboratório de Bioinformática – LABI, UNIOESTE, Foz do Iguaçu, PR
[2]Serviço de Coloproctologia, DMAD, FCM, UNICAMP, Campinas, SP

Objetivos

Realizar um estudo complementar ao apresentado em [1] e propor uma otimização do algoritmo Força Bruta (FB) para identificação de *motifs* em Séries Temporais (ST).

Métodos/Procedimentos

A identificação de padrões morfológicos (*motifs)*, mediante o FB requer grande esforço computacional, com complexidade quadrática em relação ao tamanho da ST [2]. No entanto, esse método é amplamente utilizado, pois apresenta alta precisão, e consiste em procurar subsequências de tamanho *m* em uma ST de tamanho *n*, para *m* << *n*. Para isso, cada subsequência de tamanho *m* é comparada com as subsequências restantes da ST [1]. A similaridade entre subsequências é determinada pela distância euclidiana. Para a avaliação do FB foram extraídas subséries de 1000 até 6500 com incrementos de 500 observações da ST de Eletroencefalograma (EEG)[1]. Após, para cada subsérie foram inseridas, aleatoriamente, duas ocorrências de uma subsequência com 250 observações representando um *motif* artificial. Para diminuir a casualidade, o processo de inserção foi repetido dez vezes. Os experimentos foram realizados em um computador com CPU Intel Core 2 Duo 2.20 GHz, 4 GB de RAM, sistema operacional Windows XP *Service Pack* 3 e implementados na linguagem de programação R [3]. Com o intuito de melhorar o desempenho do FB propõe-se a utilização do conceito de divisão e conquista, segmentando a busca por *motifs* em subproblemas, os quais serão resolvidos em paralelo, usufruindo com maior completude dos recursos de hardware. Para dar suporte a essa proposta serão utilizados pacotes do R tais como *multicore* e *doSMP* [3].

Resultados

No gráfico da Figura 1 é possível observar a relação existente entre o tamanho da ST e o custo em minutos para a execução do FB.

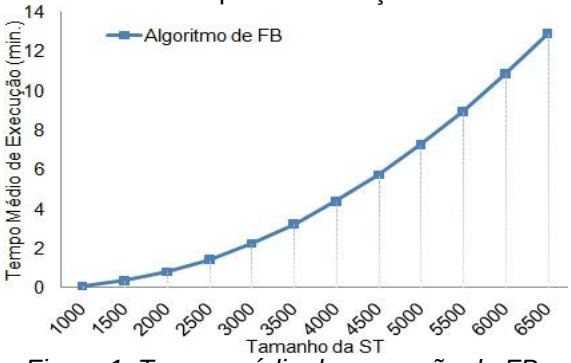

Figura 1: Tempo médio de execução do FB.

Conclusões

Embora o FB apresente elevado custo computacional, espera-se que a proposta de paralelismo permita reduzir o tempo de execução, possibilitando sua utilização em problemas reais. Como trabalhos futuros, serão desenvolvidos a proposta de paralelismo do FB e o mesmo será aplicado a dados médicos, especificamente a dados de manometria anorretal.

Referências Bibliográficas

[1] Mertz JMA, Maletzke AG, Lee HD, Wu FC. Estudo do Comportamento do Algoritmo de Força Bruta na identificação de *motifs*. In XX EAIC – Encontro Anual de Iniciação Científica, 2011. (Aceito para publicação).

[2] Chiu B, Keogh E, Lonardi S. Probabilistic discovery of time series motifs. In 9th International Conference on Knowledge Discovery and Data Mining, p.493-498, 2003.

[3] R Development Core Team. R: a language and environment for statistical computing. Disponível em: www.r-project.org. Acessado em 31 ago 2011.

[1] ST proveniente da base de dados *UCR Time Series Classification/Clustering.*