Uma Avaliação sobre a Identificação de *Motifs* em Séries Temporais*

André Gustavo Maletzke[1,2,3], **Gustavo E.A.P.A Batista**[1,3], **Huei Diana Lee**[2,3]

[1]Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo
Laboratório de Inteligência Computacional – LABIC
Caixa Postal 668, 13560-970 – São Carlos, SP, Brasil

[2]Centro de Engenharias e Ciências Exatas – Universidade Estadual do Oeste do Paraná
Laboratório de Bioinformática – LABI
Parque Tecnológico Itaipu – PTI
Caixa Postal 39, 85856-970 – Foz do Iguaçu, PR, Brasil

[3]Centro de Estudos Avançados em Segurança de Barragens – CEASB
Parque Tecnológico Itaipu – PTI

{andregm,gbatista}@icmc.usp.br, hueidianalee@gmail.com

Abstract. *In the last decade, it has increased the interest of researchers and professionals from various areas in the analysis of data that have a temporal dependency, aiming to identify patterns and relationships that might exist in data over the time. An approach recently introduced in temporal data mining is the identification of frequently occurring patterns, namely motifs, in order to describe time series. This paper presents an experimental evaluation of the process of identifying motifs, as well as a methodology for mining time series. This methodology will be applied to real data obtained from monitoring procedures related to dams' security area.*

Resumo. *Na última década tem aumentado o interesse de pesquisadores e profissionais de diversas áreas na análise de dados que possuem uma dependência temporal, buscando-se identificar padrões e relações que possam existir entre os dados ao longo do tempo. Uma abordagem recentemente introduzida na área de mineração de dados temporais consiste na identificação de padrões freqüentes, denominados motifs, para descrever séries temporais. Neste trabalho é apresentada uma avaliação experimental do processo de identificação de motifs, bem como uma metodologia para mineração de séries temporais. Essa metodologia será aplicada em dados reais obtidos a partir de processos de monitoração relacionados à área segurança de barragem.*

1. Introdução

Processos computacionais como o de Mineração de Dados — MD — [Rezende 2003, Witten and Frank 2005] são cada vez mais aplicados nas distintas áreas do conhecimento. As abordagens tradicionais de MD restringem-se a dados independentes e identicamente

*Trabalho desenvolvido com o apoio do Programa de Desenvolvimento Tecnológico Avançado — PDTA-FPTI/BR — e do Centro de Estudos Avançados em Segurança de Barragens — CEASB.

distribuídos — *i.i.d* —, portanto a ordem dos exemplos não tem significado para o processo de mineração. No entanto, em muitos domínios os dados podem possuir uma característica seqüencial e/ou temporal, por exemplo, gerados por meio de processos de monitoração realizados seqüencialmente ao longo do tempo, caracterizando um tipo de dado denominado de Série Temporal — ST — [Morettin and Toloi 2006]. Existem diversas aplicações nas quais a característica temporal e/ou seqüencial dos dados é importante no processo de mineração, pois podem representar conceitos relevantes [Last et al. 2001].

Para que o processo de mineração de dados possa ser aplicado a dados temporais, mediante técnicas tradicionais de Aprendizado de Máquina — AM —, é necessário que esses dados sejam, inicialmente, pré-processados com o intuito de extrair informações (padrões) relevantes que descrevam esses dados. Uma abordagem recentemente introduzida na área de mineração de dados temporais consiste na identificação de *motifs* para descrever uma ST. Um *motif*, também denominado de *frequent pattern*, é basicamente um padrão freqüente desconhecido dentro de uma ST, o qual possui o poder de descrever essa série.

O objetivo deste trabalho, em andamento, é apresentar e avaliar um método para a identificação de *motifs* (padrões) aplicado a séries temporais. É também apresentada, brevemente, uma metodologia aplicada à mineração de dados temporais por meio da representação de séries temporais no formato atributo-valor e posterior aplicação de métodos da aprendizado de máquina.

O restante deste trabalho está organizado do seguinte modo: na Seção 2 é apresentada uma metodologia para mineração de séries temporais por meio da extração de características e identificação de *motifs*; na Seção 3 são apresentados conceitos referentes à ST, representação de ST e um método para identificação de *motifs*. Na Seção 4 é apresentada uma avaliação do método apresentado e na Seção 5 as conclusões e trabalhos futuros.

2. Metodologia para Mineração de Dados Temporais

Ao contrário do que ocorre na maioria dos problemas de mineração de dados tradicionais, em ST a ordem dos dados é crucial para a análise. Desse modo, esses dados devem ser processados por meio de técnicas que consideram o fator temporal para que possam ser utilizados por métodos de aprendizado de máquina. Nesse contexto, é proposta uma metodologia aplicada à mineração de dados temporais baseada na representação de ST no formato atributo-valor por meio da extração de características e identificação de *motifs*. Essa metodologia está dividida em três fases:

1. **Pré-processamento de Séries Temporais**: esta fase caracteriza-se pelo pré-processamento de ST na qual métodos para remover e/ou amenizar possíveis ruídos são aplicados. Os principais problemas encontrados em uma ST referem-se à influência de tendências, dados com diferenças de escala e de amplitude e resíduos e/ou ruído branco [Shumway and Stoffer 2006]. A fase de pré-processamento de dados é considerada uma das tarefas mais trabalhosas e demoradas, sendo de fundamental importância para assegurar que os dados sejam de boa qualidade e apropriados para que possam ser analisados [Pyle 1999] e, neste contexto para que possam ser submetidos à próxima etapa da metodologia;

2. **Extração de Características e Identificação de *Motifs***: nesta fase os dados de ST pré-processados são representados no formato atributo-valor. Para isso podem ser selecionadas duas abordagens. Na primeira são definidas características a serem extraídas da ST, baseadas em medidas estatísticas. Na segunda é aplicada a identificação de possíveis *motifs* presentes em uma ST. Para tanto é necessário que uma ST passe por um processo de discretização para possibilitar a identificação dos *motifs* existentes na série. A necessidade de se discretizar essa ST deve-se ao fato que os métodos utilizados para a identificação de *motifs* são baseados na análise de dados de seqüenciamento genético e portanto recebem como entrada dados discretizados [Lin et al. 2002];
3. **Extração de Conhecimento em Base de Dados Temporais:** nesta fase a ST representada em formato estruturado é submetida, inicialmente a algoritmos tradicionais de AM com intuito de gerar um modelo que represente esses dados. Posteriormente, o conhecimento embutido nesse modelo deverá ser avaliado por meio de medidas de avaliação e, após por especialistas do domínio, os quais auxiliarão a validar tal conhecimento. Espera-se que o conhecimento encontrado, por meio dessa metodologia, possa auxiliar a especialistas no processo de tomada de decisão e/ou no direcionamento de pesquisas futuras dos mais variados domínios.

3. Processo de Identificação de *Motifs*

Uma área de estudo recente dentro da mineração de ST refere-se à mineração de padrões previamente desconhecidos. Esses padrões são sub-seqüências que ocorrem simultaneamente em um conjunto de seqüências relacionadas e, comumente, são denominados de *motifs* [Ferreira et al. 2006]. Portanto, pode-se considerar um *motif* como sendo um padrão freqüente dentro de uma determinada seqüência. Na Figura 1 é apresentada uma ST gerada artificialmente na qual foram identificados possíveis *motifs* que caracterizam essa série.

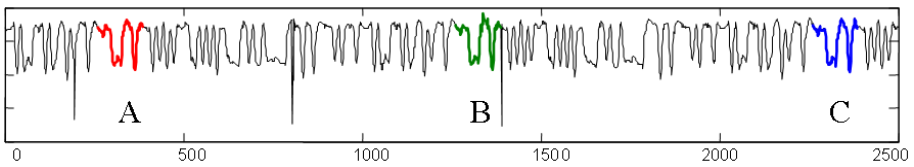


Figura 1. Exemplo da ocorrência de um *motif* em uma ST [Chiu et al. 2003].

Existem distintos métodos aplicados à identificação de *motifs*, porém para melhor entendimento do processo de identificação alguns conceitos são apresentados a seguir.

Definição 1 *(Série Temporal) [Chiu et al. 2003] Uma ST T de tamanho m é um conjunto ordenado de valores $(t_1, t_2, \ldots, t_m)$, onde $t_i \in \Re$.*

É importante ressaltar que uma ST pode conter infinitas observações equiespaçadas ou não, sendo que essas observações podem assumir valores inteiros, reais e/ou nominais.

Definição 2 *(Sub-seqüência) [Chiu et al. 2003] Dada a ST T de tamanho m, uma subseqüência C de T é uma amostra contínua de T de tamanho n, sendo $n < m$. Portanto, $C = (t_p, \ldots, t_{p+n-1})$ para $1 \leq p \leq m - n + 1$.*

Para extração de todas as sub-seqüências é aplicado o conceito de janela deslizante, o qual consiste em que dada a ST T de tamanho m e uma janela de tamanho n, essa janela é deslocada sobre a ST de modo que a cada deslocamento seja extraída uma sub-seqüência do mesmo tamanho da janela [Chiu et al. 2003].

Outro conceito relevante é o de *match* (casamento), no qual dado um número real positivo R e uma ST T contendo uma sub-seqüência C iniciando na posição p e outra sub-seqüência M na posição q, considerando a distância entre dois objetos denotada por D, tem-se que caso $D(C, M) \leq R$, então assume-se que a sub-seqüência M é similar a sub-seqüência C [Chiu et al. 2003].

Porém é necessário desconsiderar as situações denominadas de *trivial matches*, em que dadas as mesmas sub-seqüências C e M iniciando nas posição p e q, considera-se um *trivial match* se $p = q$ ou caso não exista uma sub-seqüência M' iniciando em q' tal que $D(C, M') > R$, tanto para $q < q' < p$ ou $p < q' < q$.

Como mencionado, com o intuito de identificar possíveis *motifs* é necessário que a série esteja em um formato adequado para que os distintos métodos possam ser aplicados. Desse modo, a seguir é apresentado o tratamento realizado nesta fase com intuito de adequar a ST para que possa ser submetida ao método de identificação de *motifs*.

3.1. Representação de Séries Temporais

Existem diversas abordagens para se realizar a identificação de *motifs*. A abordagem mais simples, denominada de força bruta, consiste em realizar sucessivas iterações de modo que cada possível sub-seqüência da série seja comparada com as demais existentes. Essa abordagem é custosa e somente indicada para ST de baixa dimensionalidade.

Para uma ST de dimensionalidade maior é necessário, primeiramente, que a ST esteja representada em um formato adequado, com intuito de facilitar e reduzir o espaço de busca. Desse modo, nesta metodologia foi utilizado o método *Symbolic Aggregate aproXimation* — SAX —, que consiste na representação de ST por meio da utilização de um alfabeto [Chiu et al. 2003]. No entanto, antes da aplicação do SAX sobre os dados temporais é necessário que seja realizada a redução de dimensionalidade da ST. Em [Mörchen 2006] são apresentados distintos métodos de representação de ST, os quais englobam tanto métodos de discretização quanto de redução de dimensionalidade.

Neste trabalho é utilizado o método de redução de dimensionalidade proposto em [Keogh et al. 2001] denominado de *Piecewise Aggregate Approximation* — PAA. Esse método possibilita representar uma ST em uma seqüência de dimensão reduzida. Essa redução de dimensionalidade consiste na separação de uma ST C de tamanho m em k segmentos de mesmo tamanho. A ST reduzida $\bar{C}$ pode ser representada por um vetor de tamanho k, ou seja, $\bar{C} = (\bar{c_1}, ..., \bar{c_k})$. O valor de $\bar{c_i}$ é calculado pela média dos valores presentes no i-ésimo segmento. Para cada segmento é atribuído um valor, correspondente ao cálculo da média aritmética dos valores que o compõe, que irá representar a dimensão desse segmento.

Após, é aplicado o método SAX para realizar a discretização da série. Nesse método, para cada segmento gerado por meio do PAA é atribuído um símbolo de um alfabeto α, previamente determinado. Uma tabela de pontos de corte é utilizada para determinar que símbolo será atribuído a cada segmento. Essa tabela contém números

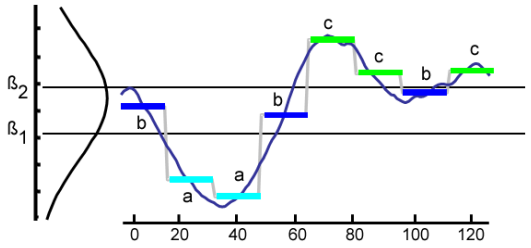
$\beta = \beta_1, ..., \beta_{\alpha-1}$, de modo que a área sob a curva Gaussiana $N(0,1)$ de β_i até β_{i-1} seja igual a $1/\alpha$, de forma a gerar símbolos equiprováveis (β_0, β_α são definidos como ∞ e $-\infty$) [Chiu et al. 2003].

Anteriormente à aplicação do PAA e SAX, é necessário que a ST seja normalizada, de modo a obter-se uma ST com média zero e desvio padrão igual a um. Na Figura 2(a) é apresentada uma tabela com os valores dos pontos de corte para alfabetos de tamanho três até seis, tais valores são provenientes de uma tabela estatística.

Na Figura 2(b) é apresentado um exemplo da aplicação de ambos os métodos PAA e SAX para um alfabeto $\alpha = \{a, b, c\}$. Valores entre β_0 e β_1 são rotulados como a, entre β_1 e β_2 como b e maiores que β_2 como c. A concatenação dos símbolos de uma sub-seqüência ou série é denominado de *word* (palavra).

β_i \ α	3	4	5	6
β_1	-0.43	-0.67	-0.84	-0.97
β_2	0.43	0	-0.25	-0.43
β_3		0.67	0.25	0
β_4			0.84	0.43
β_5				0.97

(a) Tabela com os valores dos pontos de corte para um alfabeto de tamanho de três até seis.



(b) Representação de uma ST por meio do método PAA e SAX para um alfabeto $\alpha = \{a, b, c\}$.

Figura 2. (a) Tabela com os valores dos pontos de corte e (b) representação do método SAX [Chiu et al. 2003].

Para realizar a comparação entre duas séries discretizadas é necessário definir uma distância entre dois símbolos do alfabeto. A partir da tabela de corte é construída uma matriz de distâncias $MDIST$ relacionando os símbolos do alfabeto utilizado para discretizar a série. Na Tabela 1 é apresentada a $MDIST$ construída para o alfabeto $\alpha = \{a, b, c\}$, por meio da análise da tabela de pontos de corte. Desse modo, por meio da $MDIST$ a distância de a até c, denotada por $D(a, c)$, é equivalente a $0,67$.

Tabela 1. Matriz de distâncias do alfabeto.

	a	b	c
a	0	0	0,67
b	0	0	0
c	0,67	0	0

Os valores de cada posição da $MDIST$ são calculados por meio da aplicação da Equação 1.

$$MDIST_{l,c} = \begin{cases} 0, \text{Se } |l - c| \leq 1 \\ \beta_{max(l,c)-1} - \beta_{min(l,c)}, \text{Caso Contrário} \end{cases} \quad (1)$$

Portanto, a cálculo da distância entre duas sub-seqüências discretizadas pode ser realizado, por exemplo, por meio do somatório entre a distância de cada símbolo da sub-seqüência, considerando sub-seqüências de tamanhos iguais.

3.2. Método para Identificação de *Motifs*

O método de identificação de *motifs* aplicado nesta fase é baseado no método apresentado em [Chiu et al. 2003] e consiste, basicamente, de três etapas: construção da matriz de sub-seqüências, construção da matriz de colisão e análise da matriz de colisão.

Construção da matriz de sub-seqüências

Na primeira etapa é realizada a extração de todas as sub-seqüências de tamanho n existentes na ST em questão, por meio do conceito de janela deslizante. Para cada sub-seqüência extraída é realizada a normalização e é aplicado o método de redução de dimensionalidade PAA. Após, a sub-seqüência é discretizada por meio do SAX. Em seguida, cada sub-seqüência é armazenada em uma matriz M, a qual mantém a informação de localização (eixo temporal) da sub-seqüência na ST original.

É importante ressaltar que as sub-seqüências consecutivas consideradas idênticas, após discretizadas, são descartadas, de modo que somente a primeira é mantida na matriz M. As sub-seqüências descartadas nesta etapa ainda poderão ser recuperadas em etapas posteriores, por meio da análise da lacuna existente entre a informação de localização de duas sub-seqüências consecutivas presentes na matriz M. Na Figura 3 (a) é apresentada uma representação esquemática dessa etapa utilizando uma janela $n = 16$, a qual após discretizada é representada por uma sub-seqüência (*word*) $w = 4$. Já na Figura 3 (b) é apresentada a matriz M.

Construção da matriz colisão

Após construída a matriz M é realizada a construção da matriz de colisão — Figura 3 (c). A matriz de colisão é utilizada como artifício para apontar possíveis *motifs* existentes na ST, por meio das sub-seqüências contidas na matriz M. Inicialmente, a matriz de colisão é nula e possui número de linhas e colunas iguais a quantidade de sub-seqüências existentes na matriz M, isto é, cada linha e coluna representa uma sub-seqüência da matriz M.

A matriz de colisão é construída por meio de um processo de *matching*, evitando *trivial matches*, no qual cada sub-seqüência presente na matriz M é comparada com todas as demais sub-seqüências. Para realizar a comparação entre duas sub-seqüências da matriz M é utilizada uma máscara que indica quantos e quais símbolos de cada sub-seqüência serão comparados. Para o cálculo da distância entre os símbolos das máscaras é utilizada a medida apresentada anteriormente, utilizando-se a matriz de pesos referente ao alfabeto utilizado.

Na Figura 3 (b) é apresentada uma máscara de tamanho dois localizada nas duas primeiras posições das sub-seqüências. A escolha da localização da máscara é realizada de modo aleatório, porém o tamanho da máscara é um parâmetro a ser escolhido. Neste trabalho optou-se por uma máscara de tamanho dois, similar a utilizada em trabalhos presentes na literatura [Chiu et al. 2003].

Desse modo, para as sub-seqüências cuja comparação foi considerada positiva é incrementado um contador na posição da matriz de colisão, correspondente às sub-seqüências comparadas. A partir da Figura 3 (b) pode-se observar que ocorreu o casamento entre as sub-seqüências das posições $(1, 58)$ e $(2, 985)$, de modo que na matriz de

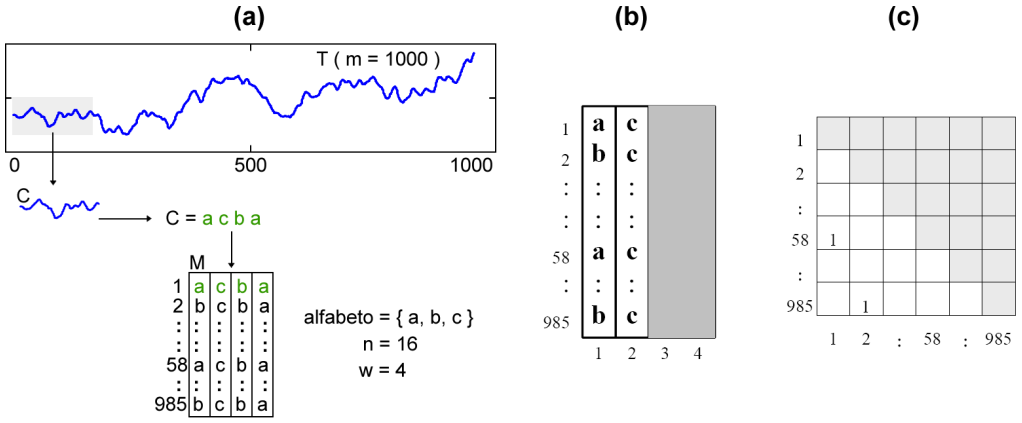


Figura 3. (a) Representação esquemática da construção da matriz M a partir de uma ST, (b) exemplo de uma máscara selecionada aleatoriamente e (c) representação da matriz de colisão [Chiu et al. 2003].

colisão da Figura 3 (c) são incrementados em um os contadores nas posições $(1, 58)$ e $(2, 985)$.

Uma única execução do processo de *matching* não é suficiente, pois as máscaras escolhidas podem ser poucos representativas. Portanto, esse processo deve ser repetido por um número adequado vezes, variando a localização da máscara. Após, é executada a terceira etapa do método.

Análise da matriz de colisão

Na terceira etapa é realizada a análise da matriz de colisão, caso as entradas da matriz sejam relativamente uniformes, isso indica que nenhum *motif* foi identificado na ST. Um valor alto em uma posição da matriz de colisão não é uma garantia da existência de um *motif*.

Para identificar um *motif* inicia-se recuperando na ST original as sub-seqüências que obtiveram valor alto na matriz de colisão. Após é calculada a distância entre essas sub-seqüências utilizando a distância Euclidiana. Considerando que duas sub-seqüências estão dentro de uma distância R estas podem ser consideradas como *motifs*. No entanto, pode haver outra sub-seqüência que também está dentro de R e precisa ser adicionada à condição de *motif*. Existem várias abordagens para identificar outras sub-seqüências que possuem potencial de *motif* [Chiu et al. 2003]. Neste trabalho foi realizada uma busca seqüencial a partir da sub-seqüência definida como *motif* por toda a ST.

4. Avaliação do Método para Identificação de *Motifs*

Com o objetivo de avaliar o método de identificação de *motifs*, e o processo como um todo, foram selecionadas três ST: $RandomWalk$, $Tide$ e $Burst$[1] [Lin et al. 2002, Chiu et al. 2003], das quais foram extraídas três sub-séries, T_1, T_2 e T_3, de tamanho 500, 1000 e 1500, respectivamente. Foram extraídos três *motifs* M_1, M_2 e M_3 de tamanhos 50, 100 e 150, representando 10% de cada sub-série, de posições aleatórias das séries $RandomWalk$, $Tide$ e $Burst$, respectivamente, com a restrição de que não sejam extraídos dos intervalos compreendidos pelas sub-séries T_1, T_2 e T_3.

[1] `http://www.cs.ucr.edu/~eamonn/time_series_data/`

Foram inseridas duas ocorrências dos *motifs* M_1, M_2 e M_3 em posições aleatórias, restringindo a ocorrência de superposição, nas sub-séries T_1, T_2 e T_3, respectivamente. Após, foi executado o método de identificação de *motifs*, apresentado neste trabalho, e verificada a eficiência na identificação dos *motifs*. Com intuito de restringir a casualidade, o processo foi repetido dez vezes para cada uma das sub-séries.

Para avaliar o desempenho do método próximo a situações reais foi realizada a inserção de um ruído Gaussiano, em cada uma das sub-séries, após a inserção dos *motifs*. Para cada sub-série foi utilizado um ruído cujo desvio padrão representa uma fração do desvio padrão da sub-série. Neste trabalho foram utilizados ruídos com um desvio padrão de 5% e 10% em relação ao desvio padrão de cada sub-série.

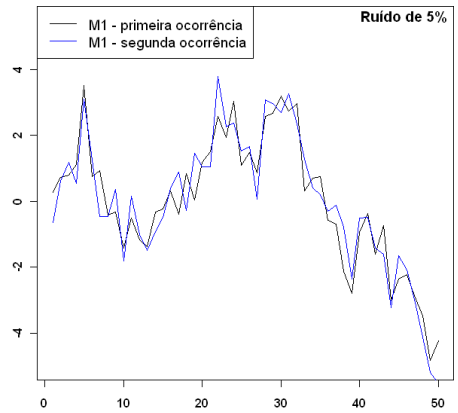
Na Tabela 2 são apresentados os resultados da execução do método para as sub-séries com e sem ruídos. Na primeira coluna são apresentadas as sub-séries T_1, T_2 e T_3, na segunda e terceira colunas são apresentados os tamanhos de cada sub-série e os tamanhos dos *motifs* inseridos. Nas colunas restantes, são apresentadas as percentagens médias de identificação dos *motifs* juntamente com o desvio padrão referentes a cada uma das sub-séries, com e sem inserção de ruído.

Tabela 2. Resultados da execução do método para identificação de *motifs*.

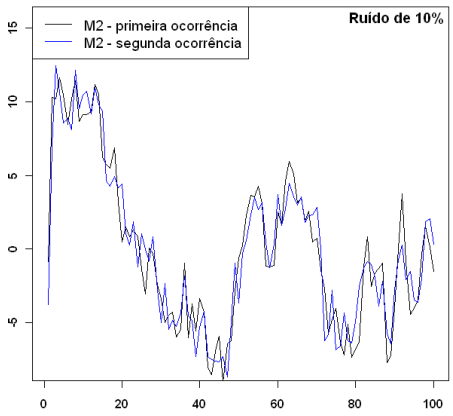
Série Temporal	Tamanho da Sub-série	Tamanho do *Motif*	Percentagem de Identificação		
			Sem Ruído (DP)	Com Ruído	
				5% (DP)	10% (DP)
T_1	500	50	100 (0.0)	90 (0.3)	100 (0.0)
T_2	1000	100	100 (0.0)	100 (0.0)	100 (0.0)
T_3	1500	150	100 (0.0)	100 (0.0)	100 (0.0)

Essa avaliação permitiu obter um melhor entendimento dos conceitos aplicados durante todo o processo de identificação de *motifs*, principalmente em relação à identificação de *motifs* que não apresentam morfologias idênticas. Na Figura 4(a) e 4(b) são apresentados gráficos que representam duas ocorrências dos *motifs* M_1 e M_2 na sub-série T_1 e T_2 em relação à inserção dos ruídos de 5% e de 10%, respectivamente. A partir desses gráficos pode-se visualizar diferenças entre as sub-seqüências que foram identificadas como *motifs* pelo método apresentado.

Observou-se que a execução do método apresentado para as sub-séries T_1, T_2 e T_3, com e sem ruídos, obteve uma taxa de identificação de 100%, exceto para T_1 com inserção de um ruído de 5%, o qual obteve uma taxa de identifiação de 90%, isto é, somente um caso não foi identificado. Entre os distintos motivos que podem estar relacionados a esse fato, está a localização na qual o *motif* M_1 foi inserido na sub-série T_1, visto que as sub-séries T_1, T_2 e T_3 apresentam características distintas, como desvio padrão e presença de sazonalidade. O caso não identificado localiza-se nas primeiras posições da sub-série T_1, as quais exibem um comportamento irregular se comparadas ao restante da sub-série, de modo que os *motifs* inseridos próximos a essas posições apresentaram maior valor de dissimilaridade. Desse modo, a aplicação de métodos de pré-processamento com o intuito de amenizar essas diferenças poderá contribuir para o processo de identificação de *motifs* em ST que apresentam essas características.



(a) Distinção entre as duas ocorrências de M_1 na sub-série T_1 com ruído de 5%.



(b) Distinção entre as duas ocorrências de M_2 na sub-série T_2 com ruído de 10%.

Figura 4. Efeito nos *motifs* M_1 e M_2 da inserção de ruído.

5. Conclusão e Trabalhos Futuros

Neste trabalho foi apresentada uma avaliação experimental da abordagem para a identificação de *motifs* apresentada em [Chiu et al. 2003], por meio da redução de dimensionalidade e discretização de ST. Essa avaliação teve o intuito de verificar a eficiência do método de identificação. A avaliação realizada possibilitou uma melhor compreensão dos conceitos e parâmetros que envolvem a aplicação do processo de identificação de *motifs* apresentado.

A identificação de *motifs* é uma das abordagens da metodologia apresentada neste trabalho. A partir dessa abordagem pretende-se obter uma representação estruturada das ST, de modo que essas séries possam também ser mineradas por algoritmos tradicionais de aprendizado de máquina. Desse modo, a avaliação apresentada constitui um estudo importante para o andamento deste trabalho e a realização de trabalhos futuros.

Embora, a avaliação realizada tenha sido bastante controlada, por meio da inserção artificial de *motifs*, a aplicação de ruído sobre as sub-séries possibilitou uma maior aproximação a casos reais. Portanto, considera-se o resultado da avaliação satisfatório. Porém, ressalta-se a necessidade de avaliações mais completas, principalmente, a aplicação em casos reais.

Como trabalhos futuros cita-se a avaliação da metodologia como um todo, por meio de *benchmarks*, a consideração de métodos para remoção de ruídos no processo de identificação de *motifs*, assim como a realização de avaliações mais completas. Por último, a aplicação da metodologia a dados relacionados ao tema segurança de barragens, obtidos da usina hidrelétrica de Itaipu, com o intuito de validar a metodologia juntamente com especialistas do domínio.

Nesse contexto, este trabalho está inserido no Projeto de Análise Inteligente de Dados no Centro de Estudos Avançados em Segurança de Barragens — CEASB — o qual está sendo desenvolvido em parceria entre o Laboratório de Inteligência Computacional — LABIC — da Universidade de São Paulo — USP — e o Laboratório de Bioinformática — LABI — da Universidade Estadual do Oeste do Paraná — UNIOESTE.

A segurança de barragens é um dos temas que tem despertado interesse e preocupação de instituições e profissionais. A avaliação de riscos de uma barragem deve permitir a identificação de problemas e auxiliar na recomendação de reparos corretivos, restrições operacionais e/ou modificações quanto às análises e aos estudos para determinar as soluções para eventuais problemas. As avaliações realizadas são baseadas, normalmente, em um conjunto de dados que é adquirido por meio de processos de monitoração e auscultação da barragem, podendo gerar um volume considerável de observações ordenadas no tempo. Nosso objetivo futuro é aplicar a metodologia proposta neste trabalho para a extração de conhecimento simbólico a partir de ST com dados de monitoração de barragens.

Referências

Chiu, B., Keogh, E., and Lonardi, S. (2003). Probabilistic discovery of time series motifs. In *Proceedings of the 9th International Conference on Knowledge Discovery and Data Mining*, pages 493–498, New York, USA. ACM Press.

Ferreira, P. G., Azevedo, P. J., Silva, C. G., and Brito, R. M. M. (2006). Mining approximate motifs in time series. In *Proceedings of the 9th International Conference on Discovery Science*, volume 4265 of *Lecture Notes in Computer Science*, pages 89–101, Barcelona, España. Springer.

Keogh, E. J., Chakrabarti, K., Pazzani, M. J., and Mehrotra, S. (2001). Dimensionality reduction for fast similarity search in large time series databases. *Knowledge and Information Systems*, 3(3):263–286.

Last, M., Klein, Y., and Kandel, A. (2001). Knowledge discovery in time series databases. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 31(1):160–169.

Lin, J., Keogh, E., Lonardi, S., and Patel, P. (2002). Finding motifs in time series. In *Proceedings of the Second Workshop on Temporal Data Mining at the Eighth International Conference on Knowledge Discovery and Data Mining*, pages 53–68, Edmonton, Alberta, Canada. ACM Press.

Morettin, P. A. and Toloi, C. M. (2006). *Análise de Séries Temporais*. Edgard Blücher, São Paulo, Brasil, 2 edition.

Mörchen, F. (2006). Time series knowledge mining. Dissertação de mestrado, Department of Mathematics and Computer Science–Philipps-University, Marburg, Hesse, Germany. Disponível em: `http://www.mybytes.de/papers/moerchen06tskm.pdf`.

Pyle, D. (1999). *Data Preparation for Data Mining*. Morgan Kaufmann, Califórnia, USA.

Rezende, S. O. (2003). *Sistemas Inteligentes: Fundamentos e Aplicações*. Manole, Barueri, Brasil.

Shumway, R. H. and Stoffer, D. S. (2006). *Time Series Analysis and its Aplications: with R examples*. Springer, New York, USA, 2 edition.

Witten, I. H. and Frank, E. (2005). *Data mining: practical machine learning tools and techniques*. Elsevier, San Francisco, California, USA, 2 edition.