Mineração de Séries Temporais por meio da Extração de Características e da Identificação de *Motifs*

André Gustavo Maletzke[1,2], Gustavo E.A.P.A Batista[1], Huei Diana Lee[2], Feng Chung Wu[2]

[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

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

Abstract. *One of the most important challenges in machine learning is the integration of temporal data to the data mining process. However, there is a lack of methods able to induce symbolic and intelligible knowledge from this data, resulting that time series data are usually treated in an adhoc manner. This work proposes a methodology to extract knowledge from time series data using characteristic extraction and motif identification. This methodology aims to induce more precise models and more comprehensible symbolic models, in particular. In addition, the methodology uses a less computationally intensive method to search for motifs. The experimental results obtained were significantly better when compared with one of the most used approaches to classify time series.*

Resumo. *Um dos grandes desafios em aprendizado de máquina é a integração de dados temporais ao processo de mineração de dados. Todavia, existe uma carência por métodos capazes de induzir conhecimento simbólico e inteligível a partir desses dados, implicando que sejam tratados de maneira adhoc. Neste trabalho é proposta uma metodologia para extração de conhecimento de séries temporais, por meio da extração de características e da identificação de motifs buscando construir modelos, principalmente simbólicos, mais precisos e compreensíveis. É utilizado também um método que demanda menor esforço computacional para a identificação de motifs. Os resultados foram significativamente melhores em relação a uma das abordagens comumente utilizada.*

1. Introdução

Atualmente, um dos grandes desafios em Aprendizado Máquina – AM – é a integração de dados temporais e sequenciais ao processo de Mineração de Dados – MD – [Yang and Wu 2006]. Existe um grande número de aplicações emergentes que envolvem o aprendizado de uma função $y = F(x)$, na qual as variáveis x e y são objetos complexos como Séries Temporais – ST. Algumas dessas aplicações incluem a identificação de transações fraudulentas em cartões de crédito e ligações telefônicas, a detecção

de intrusão em sistemas computacionais, a predição de estruturas secundárias de proteínas, a análise de dados de monitoramento de processos de manufatura, entre muitas outras [Last et al. 2001].

A maioria dos sistemas de AM não suporta esse tipo de tarefa, implicando que dados temporais e sequenciais sejam tratados de uma maneira *adhoc*. Uma das abordagens *adhoc* mais utilizadas consiste em aplicar uma janela deslizante de maneira que cada deslocamento da janela dê origem a um vetor de atributos x_i o qual deve predizer um valor individual y_i [Weiss and Indurkhya 1998]. Os pares (x_i, y_i) são fornecidos a um sistema de AM, que trata os exemplos como Independentes e Identicamente Distribuídos – i.i.d.

Obviamente, essa abordagem possui limitações. A mais evidente delas é o fato de que os pares (x_i, y_i) não são i.i.d., tampouco os atributos que constituem o vetor x_i são independentes entre si. Dessa maneira, essa abordagem claramente acarreta em perda de informação que não é utilizada no processo de indução do conhecimento, bem como não se faz presente no conhecimento induzido.

Por outro lado, independentemente dos inúmeros progressos obtidos na análise de dados sequenciais e ST nas mais diversas áreas de pesquisa, ainda existe uma carência por métodos capazes de induzir conhecimento simbólico e inteligível a partir de dados dessa natureza [Kadous and Sammut 2004]. A indução de conhecimento simbólico não se limita a fornecer predições para eventos futuros, mas também possibilita que os processos que regem a aplicação sob investigação sejam melhor compreendidos.

Neste trabalho é proposta uma nova abordagem para extração de conhecimento a partir de ST, baseada em duas estratégias: a primeira por meio de características, estatísticas globais como média, variância, mínimos e máximos globais; a segunda através de *motifs*, os quais são subsequências que se repetem em uma ST e que frequentemente representam fenômenos locais de interesse. Em conjunto, características e *motifs* fornecem atributos globais e locais que são utilizados para compor uma tabela atributo-valor que é posteriormente utilizada para a extração de conhecimento por métodos de AM.

Um grande desafio está no custo computacional da busca para identificação de *motifs*. O algoritmo força-bruta tem complexidade de tempo de $O(m^2)$, sendo m o número de observações da ST. Tal complexidade torna o algoritmo ineficiente mesmo para ST de tamanho médio. Neste trabalho é utilizado um algoritmo probabilístico para essa tarefa [Chiu et al. 2003]. Esse algoritmo permite encontrar *motifs* a um menor custo computacional, entretanto existe a probabilidade de falsos positivos e negativos.

A avaliação experimental realizada neste trabalho indica que a metodologia proposta é capaz de se desempenhar bem em problemas de classificação de ST. A abordagem proposta é superior à abordagem que fornece a ST diretamente ao sistema de AM para a maioria dos Conjuntos de Dados – CD – avaliados. É mostrado ainda que o conhecimento extraído é mais simples e inteligível do que o extraído pela abordagem comparada.

O restante deste trabalho está organizado do seguinte modo: na Seção 2 são apresentadas as principais definições e notações utilizadas neste trabalho; na Seção 3 é apresentada a metodologia proposta para mineração de ST; na Seção 4 é apresentada a configuração experimental, resultados e discussão da avaliação realizada da metodologia; Por último, na Seção 5 são apresentadas as conclusões e trabalhos futuros.

2. Definições e Notação

A análise de dados que variam ao longo do tempo é uma tarefa que tem despertado interesse em distintas áreas. Esses dados comumente são representados como séries temporais. Uma ST pode ser definida como:

Definição 1 *(Série Temporal) [Chiu et al. 2003] Uma ST Z de tamanho m é uma coleção ordenada de valores $Z = (z_1, z_2, \ldots, z_m)$ com $z_t \in \Re$.*

Na Definição 1 o tipo de dado de cada observação refere-se a dados reais ou contínuos. No entanto, diversos métodos de pré-processamento de dados transformam observações reais em valores simbólicos. Desse modo, define-se uma ST simbólica como:

Definição 2 *(Série Temporal Simbólica) Uma ST $\hat{Z}$ de tamanho m' é uma coleção ordenada de valores $\hat{Z} = (\hat{z}_1, \hat{z}_2, \ldots, \hat{z}_{m'})$ com $\hat{z}_t \in \Sigma$. Sendo Σ um alfabeto finito.*

Ainda, em algumas aplicações têm-se o interesse em estudar pequenas porções de uma ST, denominadas de subsequências:

Definição 3 *(Subsequência) [Chiu et al. 2003] Dada a ST Z de tamanho m, uma subsequência C de Z é uma amostra contínua de Z de tamanho n, sendo $n << m$. Portanto, $C = (z_p, \ldots, z_{p+n-1})$ para $1 \leq p \leq m - n + 1$.*

As subsequências podem ser extraídas por meio de uma técnica denominada de janela deslizante, cuja definição é apresentada a seguir.

Definição 4 *(Janela Deslizante) consiste em extrair todas as subsequências de tamanho n de uma ST Z de tamanho m, ou seja, como resultado obtém-se as subsequências $(z_1, \ldots, z_n), (z_2, \ldots, z_{n+1}), \ldots, (z_i, \ldots, z_{n+i-1})$, para $1 \leq i \leq m - n + 1$.*

A seguir são apresentadas algumas definições relevantes à identificação de *motifs* como a maneira de comparação de duas subsequências.

Definição 5 *(Casamento) [Chiu et al. 2003] Dado um número real positivo r e uma ST Z contendo uma subsequência C iniciando na posição p e outra M na posição q, sendo a distância entre dois objetos denotada por D, tem-se que caso $D(C, M) \leq r$, então assume-se que M é similar a C.*

Essa definição é necessária para definir o conceito de casamento trivial. Pode-se observar que os melhores casamentos para uma subsequência, além dela mesma, tendem a estarem deslocados levemente para direita ou esquerda da subsequência em questão. Desse modo, é necessário desconsiderar situações de casamento trivial.

Definição 6 *(Casamento Trivial) [Chiu et al. 2003] Dada a ST Z e as subsequências C e M iniciando nas posições p e q, considera-se um casamento trivial entre M e C se $p = q$ ou caso não exista uma subsequência M' iniciando em q' tal que $D(C, M') > r$, para $q < q' < p$ ou $p < q' < q$.*

Esses conceitos são necessários para o entendimento da metodologia, a qual é apresentada na próxima seção.

3. Metodologia para Mineração de Dados Temporais

A metodologia proposta neste trabalho é constituída de três fases que envolvem o pré-processamento da série, a extração de características e de *motifs* e, por fim, a extração do conhecimento por métodos de AM. As três fases são apresentadas a seguir.

3.1. Primeira Fase – Pré-processamento de Séries Temporais

A tarefa de pré-processamento de dados é uma das mais custosas dentro da análise de dados, pois a qualidade dos dados reflete diretamente na qualidade do conhecimento gerado a partir desses dados [Michalski et al. 1998]. Esse problema torna-se ainda mais complexo ao se trabalhar com dados temporais, pois a ordem desses dados constitui uma informação valiosa para sua compreensão.

Nesse contexto, nesta fase as ST coletadas e armazenadas em distintos formatos são preparadas para que sejam utilizadas pelos algoritmos da próxima fase. Problemas como ST com observações faltantes e observações coletadas em escalas distintas também são considerados. Outros problemas que afetam diretamente tarefas como a comparação de duas ST referem-se a ST com *offsets* diferentes no eixo y e presença de ruídos.

3.2. Segunda Fase – Extração de Características e Identificação de *Motifs*

Nesta fase é realizada a extração de características e a identificação de *motifs* de ST, de modo que, por meio dessas informações, seja possível a construção de uma tabela atributo-valor. Esta fase está dividida em duas etapas, apresentadas a seguir.

3.2.1. Etapa 1: Extração de Características

Nesta etapa são definidas as características que serão extraídas das ST, as quais são baseadas em medidas estatísticas. São utilizados os valores máximo e mínimo globais, média e variância. A partir da definição do conjunto de características, é realizada a extração dessas características produzindo um vetor de características, de modo que cada ST passe a ser representada por esse vetor.

A utilização dessas características é bastante dependente das ST que serão submetidas ao processo de extração, pois podem contribuir de maneira eficiente em casos em que o comportamento global da ST é importante para a análise. No entanto, em aplicações reais, outras características relacionadas ao domínio da aplicação podem ser consideradas. Na Figura 1 é ilustrada uma representação esquemática dessa etapa.

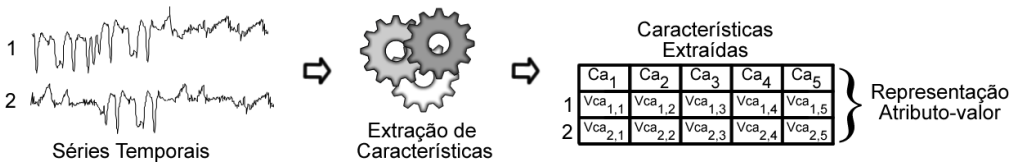


Figura 1. Representação esquemática da etapa de extração de características

3.2.2. Etapa 2: Identificação de *Motifs*

Os *motifs* existentes em uma ST são definidos como ocorrências aproximadas de uma subsequência dessa ST em posições significativamente distintas [Yankov et al. 2007]. Uma das maiores dificuldades em se identificar *motifs* é causada pela presença de casamentos triviais que fazem com que sejam identificados falsos *motifs*. Casamentos triviais são decorrentes da característica sequencial dos dados. Se uma subsequência C é comparada com as subsequências da ST Z e existir um casamento na posição p com distância $d < r$, então existe uma grande probabilidade de que as subsequências de Z nas posições ao redor de p também forneçam distâncias menores do que r, quando comparadas à C.

Como mencionado, a identificação de *motifs* por força-bruta exige uma complexidade de tempo de $O(m^2)$. Para diminuir o tempo de execução, é utilizado um método baseado na abordagem de *Random Projection* apresentada em [Chiu et al. 2003]. Em [Maletzke et al. 2008] esse processo é dividido em três passos, descritos a seguir.

Passo 1: Construção da matriz de subsequências: o processo de construção da Matriz de Subsequências – MS – consiste em extrair todas as subsequências de tamanho n da ST, por meio do conceito de janela deslizante. Essas subsequências são normalizadas para média zero e desvio padrão um. O algoritmo de *Random Projections* foi originalmente proposto para sequências de DNA e proteínas e requer como entrada uma sequência de símbolos. Para tanto, a ST é transformada para uma série simbólica por meio do método *Symbolic Aggregate aproXimation* – SAX – [Lin et al. 2003]. Esse método realiza, em uma etapa preliminar, a redução de dimensionalidade por meio da definição de uma janela de redução. Para isso, foi utilizada uma janela correspondente a 10% do tamanho da sequência a ser discretizada. O SAX foi aplicado utilizando um alfabeto de tamanho seis, corroborando com estudos presentes na literatura. No exemplo da Figura 2 (a) são extraídas subsequências de tamanho 16 que após discretizadas formam sequências simbólicas de tamanho $n_{sax} = 4$, por meio do alfabeto (a, b, c).

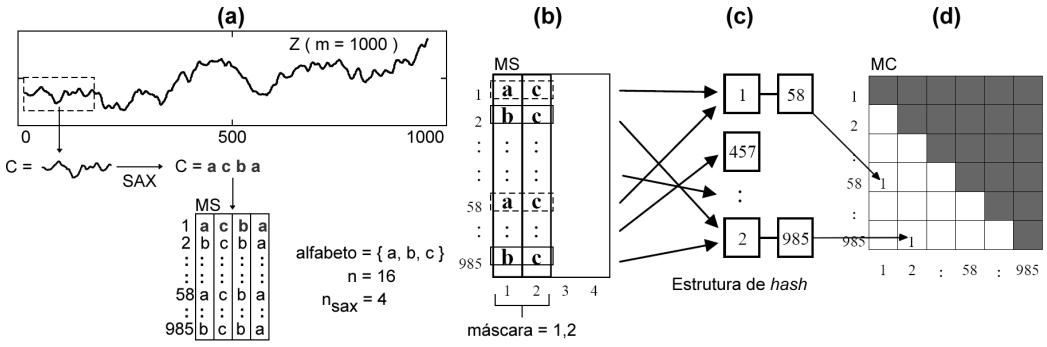


Figura 2. Representação esquemática do processo de identificação de *motifs*

Passo 2: Construção da matriz de colisão: a Matriz de Colisão – MC – é utilizada como artifício para apontar possíveis *motifs* existentes na ST. Essa matriz possui número de linhas e colunas iguais ao número de linhas da MS e, inicialmente, a MC é nula. Essa matriz é preenchida por meio de um processo iterativo, sendo que a cada iteração toda a MS é percorrida, de modo que a localização de cada subsequência seja inserida em uma estrutura *hash* – Figura 2 (c). Para isso, é utilizada uma máscara, gerada aleatoriamente, que indica quais colunas da MS devem ser utilizadas como parâmetro para a função de *hash*. Na Figura 2 (b) a máscara é igual a (1,2), portanto as subsequências das linhas (1,58) e (2, 985) da MS irão colidir ao se aplicar a função *hash*, pois ambas possuem os mesmos valores nas colunas 1 e 2. Neste trabalho é utilizada uma máscara de tamanho dois, conforme sugerido na literatura. Ao final de cada iteração é atualizada a MC por meio do incremento de um contador nas posições correspondentes às subsequências que colidiram – Figura 2 (d). O processo é repetido um número determinado de vezes, pois as máscaras escolhidas podem ser pouco representativas. É importante ressaltar que a cada iteração a estrutura *hash* deve ser esvaziada. Neste trabalho,

as iterações realizadas são equivalentes a 50% de todas as possíveis combinações de máscaras.

Passo 3: Análise da matriz de colisão: um valor alto em uma posição da MC é um indício, embora não seja uma garantia, da existência de um *motif*. Para identificar um *motif* verifica-se na MC a localização das subsequências que obtiveram maior número de colisões. Após, é calculada a distância entre essas subsequências utilizando uma medida de similaridade. Caso duas subsequências estejam dentro de uma distância r, essas podem ser consideradas como *motifs*. Outras subsequências também podem estar dentro de r e precisam ser adicionadas à condição de *motif*. Existem várias abordagens para identificar tal situação [Chiu et al. 2003]. Neste trabalho foi realizada uma busca sequencial a partir da subsequência definida como *motif* por toda a ST. Foi utilizado um limiar (r) igual a 10%, isto é, uma percentagem que corresponde ao erro médio aceito para a diferença existente entre duas observações de duas subsequências distintas. Para o cálculo de similaridade foi utilizada a distância Euclidiana.

O processo de identificação de *motifs* utilizado é iterativo e probabilístico. Portanto, devido a essa característica e a de não explorar todo o espaço de busca, esse processo é mais eficiente em relação ao esforço computacional se comparado com a abordagem força-bruta [Chiu et al. 2003].

Na Figura 3 é ilustrado esquematicamente o processo de mapeamento de um conjunto de ST para essa representação por meio de quatro *motifs*, previamente identificados.

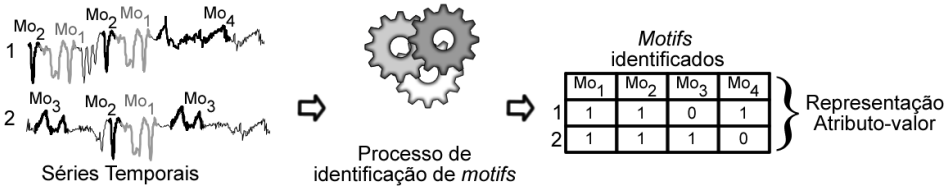


Figura 3. Representação esquemática da etapa de identificação de *motifs*

Desse modo, os *motifs* identificados juntamente com as características podem ser representados em uma tabela atributo-valor em que os atributos referem-se às características extraídas e à presença (0 ou 1) dos *motifs* identificados. Neste trabalho, foram identificados *motifs* de distintos tamanhos, de acordo com um percentual do tamanho de cada ST. Para tanto, foi utilizada uma faixa de 5% até 30%, com incrementos de 5%. A necessidade de se procurar por *motifs* de diversos tamanhos é decorrente do fato que não se conhece, a priori, quais os *motifs* existentes nas ST.

Essa metodologia foi desenvolvida utilizando tecnologias e ferramentas livres e implementada no ambiente estatístico e gráfico R[1].

3.3. Terceira Fase – Extração de Conhecimento em Bases de Séries Temporais

Nesta fase são aplicados métodos de AM com intuito de extrair conhecimento da tabela atributo-valor construída a partir das ST. É importante ressaltar que nesta etapa podem ser utilizados quaisquer métodos de AM que aceitem como entrada uma tabela atributo-valor.

[1] `http://www.r-project.org/`

Neste trabalho, é dado enfoque aos métodos simbólicos, pois esses métodos possibilitam a indução de modelos nos quais é possível compreender o conhecimento induzido. Como é discutido na próxima seção, os modelos simbólicos induzidos a partir da metodologia proposta tendem a ser simples e de fácil compreensão.

4. Análise Experimental

Nesta seção são apresentados os resultados da análise experimental realizada para avaliar o desempenho da metodologia proposta. Inicialmente é descrita a configuração experimental e após são apresentados os resultados e a discussão.

4.1. Configuração Experimental

A metodologia proposta foi avaliada experimentalmente em comparação com a abordagem na qual as ST são fornecidas diretamente para o algoritmo de aprendizado, chamada neste trabalho de abordagem tradicional. Essa abordagem simples é amplamente utilizada na área, sobretudo para a comparação de resultados experimentais. Para alguns CD ela é capaz de prover excelentes resultados em termos de erro de classificação. Em especial, o algoritmo k-Vizinhos mais Próximos – *kNN* – com $k = 1$ com a distância Euclidiana é bastante utilizado em classificação de ST.

A indução de modelos foi realizada por meio do aplicativo WEKA[2] e foram utilizados os algoritmos *J*48 e *kNN*, com suas configurações padrão. Foram selecionados sete CD, descritos Tabela 1.

Tabela 1. Descrição dos conjuntos de dados utilizados

Conjunto de dados	# Ex.	#Observações da ST	Classes	% Classe	Erro majoritário	Valores desconhecidos
ECG	200	96	1 2	66,5% 33,5%	33,5%	Não
FaceFour	112	350	1 2 3 4	19,6% 30,4% 25,9% 24,1%	69,6%	Não
Coffee	56	286	1 2	48,2% 51,8%	48,2%	Não
Beef	60	470	1 2 3 4 5	20,0% 20,0% 20,0% 20,0% 20,0%	80,0%	Não
Trace	200	275	1 2 3 4	25,0% 25,0% 25,0% 25,0%	75,0%	Não
Wafer	7164	128	1 2	10,6% 89,4%	10,6%	Não
Gun-Point	200	150	1 2	50,0% 50,0%	50,0%	Não

Como medida de avaliação foi utilizada taxa média de erro estimada por meio do método de amostragem 5×2 *fold cross-validation*. Optou-se pela taxa média de erro, pois é amplamente utilizada na área de classificação de ST. A abordagem de amostragem 5×2 *fold cross-validation* foi escolhida uma vez que a maioria dos CD selecionados possuem um número restrito de exemplos, sendo que cada exemplo corresponde a uma ST.

[2] `http://www.cs.waikato.ac.nz/ml/weka/`

Os CD utilizados foram obtidos do repositório de dados da UCR *Time Series Classification/Clustering*[3]. Todos os CD utilizados são amplamente utilizados por trabalhos da literatura e de acesso público, com o objetivo de facilitar a reprodução dos resultados e a comparação com métodos propostos por outros pesquisadores.

A avaliação dos resultados foi realizada por meio do teste estatístico t-*Student*, utilizando o ambiente estatístico e gráfico R.

4.2. Resultados e Discussão

Na Tabela 2 são apresentadas as taxas médias de erro e os respectivos desvios padrão dos classificadores induzidos para cada CD. As menores taxas são apresentadas em negrito e as que apresentaram Diferença Estatísticamente Significativa – **d.e.s** – são assinaladas na coluna **d.e.s**.

Nessa tabela pode-se observar que para o *J*48 a utilização da metodologia possibilitou a indução de modelos simbólicos mais precisos, com **d.e.s** para cinco dos sete CD utilizados na avaliação. A classificação por meio do *kNN* obteve uma taxa média de erro significativamente menor para quatro CD, sendo que em um caso a utilização da abordagem tradicional apresentou uma taxa média de erro, com **d.e.s**, menor.

Tabela 2. Taxas médias de erro e respectivos desvios padrão dos classificadores

Conjunto de dados	*J*48			*kNN*		
	Metodologia proposta	Abordagem tradicional	**d.e.s**	Metodologia proposta	Abordagem tradicional	**d.e.s**
ECG	**1,00** (1,58)	23,11(3,76)	✓	26,28 (3,40)	**11,60** (3,18)	✓
FaceFour	**13,38** (5,28)	19,78 (5,65)	✓	**3,74** (2,45)	6,98 (3,71)	✓
Coffee	**9,11** (6,02)	38,96 (10,63)	✓	**7,78** (4,57)	33,66 (10,34)	✓
Beef	**43,66** (14,36)	49,67 (9,74)		**47,00** (9,74)	53,33 (11,76)	
Trace	**1,10** (1,45)	22,70 (5,14)	✓	**0,00** (0,00)	13,80 (3,43)	✓
Wafer	**0,23** (0,21)	1,05 (0,40)	✓	**0,14** (0,06)	0,23 (0,08)	✓
Gun-Point	**11,10** (5,80)	12,20 (6,61)		7,20 (6,07)	**6,60** (1,43)	

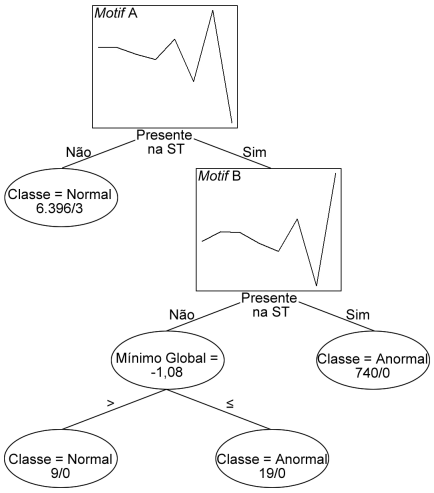
O fato do *kNN* ter apresentado **d.e.s** para um número menor de CD pode ser atribuído às características desse método. O *kNN* atribui igual peso a todos os atributos do CD, portanto, caso exista um pequeno subconjunto de atributos relevantes, esses podem ter pouca influência no processo de classificação. Nesse sentido, a metodologia proposta, geralmente, resulta em um elevado número de *motifs* para cada CD, influenciando negativamente no desempenho do *kNN*. Já o *J*48 realiza a seleção dos atributos mais relevantes, portanto, mesmo que existam atributos pouco relevantes, esses tendem a não serem considerados pelo algoritmo. Desse modo, a utilização de métodos de seleção de atributos pode auxiliar na redução da taxa média de erro obtida pelo *kNN*. Essa tarefa constitui um das atividades futuras deste trabalho.

A metodologia proposta quando comparada com a abordagem tradicional apresenta menores taxas médias de erro na maioria dos CD utilizados, sendo que para alguns casos essa diferença foi considerada alta, como nos conjuntos *Trace*, *Coffee* e ECG, para o *J*48 e para *kNN* nos conjuntos *Coffee* e *Trace*. Embora, não foi possível observar **d.e.s** em todos os casos, ocorreu uma redução na taxa média de erro na maioria dos casos. Entretanto, é importante observar que em alguns CD o desvio padrão observado foi alto, principalmente para o CD *Beef*, isso pode ser explicado pelo número reduzido

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

de exemplos e elevado número de classes desse CD. De modo geral, esses resultados são animadores, pois elucidam que a metodologia pode apresentar resultados bastante competitivos.

No caso dos modelos simbólicos, foi possível construir árvores de decisão relacionando características e *motifs*, os quais em casos reais podem trazer informações relevantes para especialistas de distintos domínios. Na Figura 4 são apresentadas as árvores construídas com a aplicação da metodologia (esquerda) e de modo parcial, devido ao tamanho, a construída pela abordagem tradicional (direita) para o CD *Wafer*. Os valores que estão entre parênteses indicam a cobertura de cada ramo até o nó folha.




```
ST[18] ≤ 1,91
|  ST[89] ≤ 0,96
|  |  ST[59] ≤ -1,07
|  |  |  ST[109] ≤ -0,41: 1 (25)
|  |  |  ST[109] > -0,41: -1 (121/1)
|  |  ST[59] > -1,07
                ⋮      ⋮
|  |  ST[108] > 0,38
|  |  |  ST[107] ≤ 0,40: -1 (4/0)
|  |  |  ST[107] > 0,40: 1 (12/0)
|  ST[129] > 0,56
|  |  ST[16] ≤ 1,06: 1 (15/0)
|  |  ST[16] > 1,06: -1 (5/0)
```

Figura 4. Árvores construída com a metodologia proposta (esquerda) e com a abordagem tradicional (direita)

O conhecimento representado por meio da árvore da Figura 4 (esquerda) é mais intuitivo e compreensível se comparado ao conhecimento representado na árvore da Figura 4 (direita), isto é, nesse exemplo o atributo referente ao *motif* A possui a capacidade de descrever grande parte dos exemplos, sendo que na abordagem tradicional a árvore gerada é complexa e são necessários vários atributos, onde cada atributo é uma observação da ST, para realizar a classificação de um exemplo. Nesse sentido, este trabalho permitiu a construção de modelos de fácil interpretabilidade, isso se deve a utilização de características de baixa complexidade e a utilização de *motifs* como atributos para a indução de modelos. É importante ressaltar que os *motifs* podem refletir acontecimentos com maior grau de detalhamento, podendo indicar comportamentos que evidenciem a presença prematura de determinados fenômenos.

De modo geral, foi observado que a maioria dos modelos simbólicos construídos por meio da metodologia apresentaram complexidade sintática menor, com **d.e.s**, em comparação com os construídos pela abordagem tradicional. Por motivos de espaço, esses resultados não são detalhados neste trabalho.

5. Conclusão e Trabalhos Futuros

De acordo com os resultados da avaliação experimental realizada foi possível observar que a metodologia proposta apresentou, para o indutor $J48$, taxas médias de erro menores em todos os CD utilizados. Para o método kNN a metodologia também obteve menor taxa

média de erro em cinco dos sete CD. Os resultados ilustram a contribuição deste trabalho, por meio do potencial da metodologia proposta na mineração de ST, especificamente na tarefa de classificação.

Como trabalhos futuros pretende-se melhorar o método de identificação de *motifs*, possibilitando a identificação de *motifs* de distintos tamanhos e a utilização de outras características que possam contribuir ainda mais na construção de modelos mais precisos, mantendo a interpretabilidade, quando aplicados em casos reais. Embora a tarefa de seleção de atributos seja amplamente utilizada em MD, esta não foi aplicada neste trabalho, pois buscou-se avaliar a metodologia de modo geral. Trabalhos futuros irão investigar o desempenho do kNN aliado a métodos de seleção de atributos. Por último, a utilização dessa metodologia em CD de ST reais, como da área de medicina e de monitoramento ambiental, nas quais estamos trabalhando juntamente com especialistas dessas áreas.

Agradecimentos: trabalho desenvolvido com o apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq, da Fundação de Amparo à Pesquisa do Estado de São Paulo – FAPESP, do Programa de Desenvolvimento Tecnológico Avançado – PDTA-FPTI/BR e do Centro de Estudos Avançados em Segurança de Barragens – CEASB.

Referências

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

Kadous, M. W. and Sammut, C. (2004). "Constructive induction for classifying time series". In: 15th European Conference on Machine Learning, volume 3201 of *Lecture Notes in Computer Science*, pages 192–204, Pisa, Italy. Springer.

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 Chiu, B. (2003). "A symbolic representation of time series, with implications for streaming algorithms". In: 8th Workshop on Research Issues in Data Mining and Knowledge Discovery, pages 2–11, San Diego, USA. ACM Press.

Maletzke, A. G., Batista, G. E., and Lee, H. D. (2008). "Uma avaliação sobre a identificação de motifs em séries temporais". In: 3th Congresso da Academia Trinacional de Ciências, volume 1, pages 1–10, Foz do Iguaçu, Brasil.

Michalski, R. S., Bratko, I., and Kubat, M. (1998). *Machine learning and data mining*. Wiley, Chichester, West Sussex, England.

Weiss, S. M. and Indurkhya, N. (1998). *Predictive Data Mining: a practical guide*. Morgan Kaufmann, California, USA.

Yang, Q. and Wu, X. (2006). "10 challenging problems in data mining research". *International Journal of information Technology and Decision Making*, 5(4):401–420.

Yankov, D., Keogh, E., Medina, J., Chiu, B., and Zordan, V. (2007). "Detecting time series motifs under uniform scaling". In: 13th International Conference on Knowledge Discovery and Data Mining, pages 844–853, New York, USA. ACM Press.