Extração de Padrões e Construção de Modelos Simbólicos para Previsão de Dados Temporais*

Carlos Andres Ferrero[1], André Gustavo Maletzke[1], Huei Diana Lee[1], Gustavo E. A. P. A. Batista[2], Wu Feng Chung[1,3]

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

[2]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

[3]Faculdade de Ciências Médicas – Universidade Estadual de Campinas
Serviço de Coloproctologia – Departamento de Moléstias do Aparelho Digestivo – DMAD
Caixa Postal 6111, 13083-970 – Campinas, SP, Brasil

{anfer86,andregustavom,hueidianalee,wufengchung}@gmail.com

Resumo. *A previsão de dados temporais é uma tarefa de crescente interesse em distintas áreas. De modo geral, os métodos propostos para tratar esse problema consistem na construção de modelos matemáticos que não consideram padrões locais contidos nos dados. Neste trabalho, é proposto um método de previsão de dados temporais que permite extrair e relacionar padrões, e utilizá-los na construção de modelos de previsão de comportamentos futuros. A previsão é tratada como um problema de classificação, permitindo a aplicação de algoritmos de aprendizado de máquina simbólico. Foi realizado um experimento preliminar utilizando dados artificiais, o qual permitiu evidenciar os relacionamentos entre as informações do passado e os comportamentos futuros.*

1. Introdução

A coleta e análise de fenômenos temporais são tarefas de crescente interesse em distintas áreas, motivada pela premissa de que o comportamento passado de um fenômeno pode influenciar no comportamento atual e futuro dos dados. Diversas tarefas de interesse estão associadas a essa analise, dentre as quais, a previsão de dados temporais consiste em uma das de maior interesse, pois permite prever dados desconhecidos a partir de um conjunto de informações conhecidas. Para isso, têm sido propostos métodos para a previsão de comportamentos lineares e não-lineares. Os primeiros, assumem que os dados respeitam alguma distribuição estatística e, com base nessa informação, é realizado o ajuste de um modelo aos dados. Porém, parte dos fenômenos naturais avaliados no tempo envolvem comportamentos não-lineares. Assim, abordagens para modelagem não-linear, também denominadas de regressões não-paramétricas, têm sido propostas [McNames 1999].

As abordagens não-lineares são comumente classificadas como globais e locais [Karunasinghe and Liong 2006]. Os métodos globais utilizam a série de dados tem-

*Trabalho realizado com auxílio do Programa de Desenvolvimento Tecnológico Avançado — PDTA/FPTI-BR — e do Conselho Nacional de Desenvolvimento Científico e Tecnológico — CNPq.

porais para a construção de um único modelo que represente todo o comportamento dos dados e os métodos locais utilizam uma parcela dessa informação considerada de maior importância para a construção do modelo. Em muitos domínios o comportamento global de dados temporais possui uma baixa capacidade de descrevê-los, assim sendo, a identificação de comportamentos locais atrelada ao comportamento global pode constituir uma informação de grande relevância para muitas tarefas, inclusive a de previsão.

Diversos estudos têm procurado identificar e descrever comportamentos locais presentes em séries de dados temporais de modo que tais comportamentos possam auxiliar na construção de modelos preditivos, descritivos e de classificação de dados temporais [Antunes and Oliveira 2001, Lin et al. 2002, Mörchen 2006, Maletzke et al. 2009].

Nesse contexto, algoritmos de Aprendizado de Máquina para construção de modelos, a partir de dados não-temporais, têm sido adaptados de modo que possam ser aplicados sobre dados temporais. Esses algoritmos, analogamente aos métodos de previsão, também realizam a inferência de um modelo baseado em experiências anteriores, porém, sem considerar a relação temporal existente nos fatos passados. Desse modo, a adaptação desses algoritmos para análise e construção de modelos a partir de dados temporais constitui uma área de estudo de crescente interesse [McNames 1999, Povinelli and Feng 2003, Kulesh et al. 2008, Ferrero et al. 2009].

Nesse cenário, o objetivo deste trabalho em andamento consiste na proposta de um novo método para previsão de comportamentos temporais não-lineares, o qual consiste na identificação de relacionamentos do tipo causa-efeito, por meio da extração de padrões, com o intuito de auxiliar na previsão de comportamentos futuros. O trabalho está inserido no projeto Análise Inteligente de Dados, desenvolvido em uma parceria 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 —, e o Serviço de Coloproctologia da Universidade Estadual de Campinas — UNICAMP / Campinas.

O restante deste trabalho está organizado da seguinte forma: na Seção 2 é descrito o método proposto neste trabalho; na Seção 3 é apresentado um experimento preliminar realizado com o intuito de ilustrar o funcionamento do método proposto; na Seção 4 são apresentadas as conclusões e os trabalhos futuros deste trabalho.

2. Método Proposto

O método proposto neste trabalho consiste na representação do problema de previsão de comportamentos futuros em dados temporais como um problema de classificação. Assim, esse método é constituído de cinco etapas: (1) definição de parâmetros iniciais; (2) determinação das classes; (3) extração de características e padrões do passado; (4) construção de modelos; e (5) avaliação dos modelos. Cada uma das etapas é descrita a seguir.

Etapa 1 — Definição de Parâmetros Iniciais

Nessa primeira etapa são definidos os parâmetros iniciais necessários para a execução das próximas etapas. Considerando a série temporal $X = (x_1, x_2, \ldots, x_n)$ de tamanho n, em que cada elemento x_t da série indica o valor amostrado no tempo t, os parâmetros a serem

determinados são: o tamanho da janela, o horizonte de previsão e o passado considerado, e são descritos a seguir:

Tamanho da janela: consiste em um w, tal que, a partir das observações de X, representadas no espaço w-dimensional, é possível identificar e modelar a trajetória da série ao longo do tempo [Chun-Hua and Xin-Bao 2004]. Essa janela é dependente da frequência de amostragem, bem como do domínio e do fenômeno avaliado. Porém, algumas abordagens permitem identificar aproximações para determinar o tamanho da janela, entre essas: pela análise visual do especialista do domínio; pelo cálculo da dimensão de correlação; e pelo método de Falsos Vizinhos Próximos (*FNN— False Nearest Neighbor*). A primeira abordagem consiste na identificação visual da periodicidade dos dados [Kulesh et al. 2008]. A segunda abordagem permite identificar a dimensão de imersão pela avaliação da auto-similaridade entre os pontos [Grassberger and Procaccia 1983]. E a última abordagem, consiste em definir w de modo que valores de dimensão superiores a w não influenciem na evolução da trajetória da série [Kennel et al. 1992, Chun-Hua and Xin-Bao 2004].

Horizonte de previsão: indica a quantidade de valores futuros, h, que serão previstos a partir de um instante t_a tal que $1 \leq t_a \leq n$. É importante ressaltar que, assim como na determinação do tamanho da janela, o horizonte de previsão também é dependente do domínio de aplicação, de acordo com o comportamento do fenômeno avaliado e da frequência de coleta das informações ao longo do tempo, bem como das questões operacionais associadas ao processo de tomada de decisão considerando um horizonte de previsão específico.

Passado considerado: define a quantidade p de valores do passado a serem considerados para a previsão do horizonte de previsão h. Como a série de valores do passado representa o conhecimento disponível para prever h, torna-se necessário utilizar um passado suficientemente longo, de modo a extrair informações representativas desses dados, que possibilitem prever, por meio de um modelo objetivo, o comportamento futuro.

Nessa fase, com base nesses parâmetros, a série temporal X é transformada para um formato estruturado. Assim, é definido o conjunto de dados $XD = \{xd_1, xd_2, \ldots, xd_m\}$, em que xd_i representa o i-ésimo exemplo e m a cardinalidade do conjunto, dada por $\lfloor t_a/w \rfloor - 1$, considerando o tempo atual t_a. Cada elemento xd_i consiste em um par ordenado $(\boldsymbol{xp_i}, \boldsymbol{xh_i})$, em que são representados os p valores do passado e os h valores futuros, respectivamente, a partir de t_a. Desse modo, o conjunto XD é definido conforme a Equação 1.

$$
\begin{aligned}
xd_1 &= ((x_{t_a-w\times 1-p}, \ldots, x_{t_a-w\times 1-1}, x_{t_a-w\times 1}), (x_{t_a-w\times 1}, x_{t_a-w\times 1+1}, \ldots, x_{t_a-w\times 1+h})) \\
xd_2 &= ((x_{t_a-w\times 2-p}, \ldots, x_{t_a-w\times 2-1}, x_{t_a-w\times 2}), (x_{t_a-w\times 2}, x_{t_a-w\times 2+1}, \ldots, x_{t_a-w\times 2+h})) \\
&\vdots \\
xd_m &= ((x_{t_a-w\times m-p}, \ldots, x_{t_a-w\times m-1}, x_{t_a-w\times m}), (x_{t_a-w\times m}, x_{t_a-w\times m+1}, \ldots, x_{t_a-w\times m+h}))
\end{aligned}
\tag{1}
$$

De acordo com essa representação, o problema de previsão consiste em criar uma hipótese $hip(\boldsymbol{xp_i})$ que seja uma aproximação da função $f(\boldsymbol{xp_i}) = \boldsymbol{xh_i}$, *i.e.*, criar um modelo que permita prever o comportamento futuro $\boldsymbol{xh_i}$ a partir do passado $\boldsymbol{xp_i}$.

Etapa 2 — Determinação das Classes

Esta etapa tem como objetivo o agrupamento dos horizontes de previsão. Para tanto, inicialmente, é construído o conjunto de horizontes $XH = \{\boldsymbol{xh_1}, \boldsymbol{xh_2}, \ldots, \boldsymbol{xh_m}\}$, em que cada elemento $\boldsymbol{xh_i}$ corresponde ao horizonte de previsão do exemplo $xd_i \in XD$. A partir do conjunto de séries XH são identificados padrões de acordo com a similaridade entre essas séries. Com isso, é definido o conjunto $C = \{c_1, c_2, \ldots, c_p\}$, contendo as p classes identificadas a partir de XH, tal que $|C| \leq m$, de modo que cada série $\boldsymbol{xh_i} \in XH$ possa ser rotulada com alguma classe contida em C. Essa identificação deve ser realizada juntamente com especialistas do domínio de aplicação. No entanto, técnicas computacionais podem dar apoio a essa tarefa, como por exemplo, pela utilização de algoritmos de agrupamento *(clustering*), que têm como finalidade construir grupos (*clusters*), a partir do conjunto de séries, de acordo com a similaridade entre as séries. Posteriormente, os agrupamentos encontrados são analisados e validados pelos especialistas do domínio.

De acordo com o modo como são definidos os agrupamentos, esses algoritmos podem ser classificados em particionais e hierárquicos. Os primeiros consistem, basicamente, em dividir o conjunto de exemplos em k partições, de modo iterativo, até que cada exemplo esteja mais próximo ao centroide do agrupamento a que pertence do que ao centroide dos outros agrupamentos [Alpaydin 2004]. Por outro lado, os algoritmos de agrupamento hierárquico possuem como característica a organização dos grupos por meio de uma estrutura hierárquica, que descreve diferentes agrupamentos a cada nível da hierarquia. Nessa última abordagem não é necessário definir, a priori, o número de agrupamentos, o que permite maior flexibilidade durante a análise dos dados, permitindo considerar diferentes graus de granularidade [Everitt 1993]. Os principais algoritmos de agrupamento hierárquico presentes na literatura são: *single-link*, *complete-link* e *average-link*. Outra questão a ser considerada é a medida de similaridade, a qual define o critério que determina, objetivamente, a semelhança entre as séries do conjunto XH. Essa medida pode apresentar forte influência no resultado da aplicação do algoritmo de agrupamento. Em [Everitt 1993] é apresentada uma descrição de várias medidas de similaridade propostas na literatura.

No final dessa etapa, cada exemplo contido em $\boldsymbol{xh_i}$ é rotulado com base nas classes, contidas em C, identificadas. Esse mapeamento permite a representação do problema de previsão de dados temporais em um problema de classificação.

Etapa 3 — Extração de Características e Padrões do Passado

Nesta etapa busca-se representar as séries xp_i no formato atributo-valor. Para tanto, é construído um novo conjunto $XD' = \{xd'_1, xd'_2, \ldots, xd'_m\}$, em que cada elemento $\boldsymbol{xd'_i}$ é um par ordenado $(\boldsymbol{xp_i}, ch_i)$, na qual o primeiro elemento corresponde ao passado do exemplo $xd_i \in XD$ e o segundo elemento corresponde à classe ch_i do horizonte de previsão $\boldsymbol{xh_i}$ identificada na etapa anterior. Com base nas m séries, são extraídos atributos, os quais consistem em medidas de estatística descritiva e padrões morfológicos. Esta fase é baseada no método proposto por [Maletzke et al. 2009] utilizado para a extração

de conhecimento em bases de dados de séries temporais por meio da combinação de duas abordagens, apresentadas a seguir:

Extração de medidas estatísticas: nesta abordagem são definidas as medidas de estatística descritiva que serão determinadas a partir das m séries. A determinação de medidas de estatística descritiva é uma abordagem bastante utilizada na representação de dados temporais e, geralmente, buscam descrever o comportamento global das séries. Distintas medidas podem ser utilizadas, desde estatística descritiva como média (f_μ), máximo (f_{max}) e mínimo (f_{min}) globais, bem como medidas definidas com o auxílio de especialistas, as quais podem ser de grande valor para o entendimento das séries. Desse modo, ao final dessa etapa cada uma das m séries temporais é representada não mais por suas observações, mas em função das medidas selecionadas;

Identificação de *motifs*: nesta abordagem é realizada a busca por sequências dentro das séries, que apresentam o mesmo comportamento morfológico, conhecidas na literatura como *motifs* [Antunes and Oliveira 2001, Lin et al. 2002]. Esses *motifs* podem constituir informações de alta relevância para o entendimento das séries temporais. No entanto, o processo de identificação desses fenômenos locais é uma tarefa custosa em relação ao esforço computacional, podendo se tornar inviável em algumas situações [Lin et al. 2002]. No intuito de contornar esse problema, é utilizado um método probabilístico baseado no método proposto em [Buhler and Tompa 2002] e adaptado por [Chiu et al. 2003] para o domínio de séries temporais. A aplicação desse método requer a definição de alguns parâmetros, dentre os quais o tamanho do *motif* que será buscado na série, definido por α, e uma medida de similaridade utilizada como critério de similaridade. A partir da definição desses parâmetros o método de identificação de *motifs* é aplicado sobre as m séries temporais agrupadas de acordo com as classes identificadas. Posteriormente, cada *motif* identificado receberá um identificador e constituirá um dos possíveis valores de cada localização temporal do passado. Desse modo, a função $f_{mo}(\boldsymbol{xp_i}, j)$ permite verificar a existência de *motifs*, dado um exemplo $\boldsymbol{xp_i}$ e uma posição j, para $(\alpha - 1) \leq j \leq p$.

Desse modo, espera-se que ao término dessa fase obtenha-se uma representação atributo-valor das m séries temporais em função de medidas de estatística descritiva e de *motifs*, descrevendo as séries por meio de características globais e locais, respectivamente. Na Tabela 1 essa ideia é ilustrada, em que as linhas representam os exemplos e as colunas os atributos. Os atributos são constituídos pelas medidas estatísticas extraídas, pelos *motifs* identificados em cada instante de tempo do passado e pela classe do comportamento futuro de cada exemplo.

Tabela 1. Tabela atributo-valor resultante da Etapa 3.

	Medidas			***Motifs***				
i	$\boldsymbol{\mu}$	$\boldsymbol{max}$	...	$\boldsymbol{p}$	...	$\boldsymbol{\alpha}$	$\boldsymbol{\alpha - 1}$	$\boldsymbol{ch_i}$
1	$f_\mu(\boldsymbol{xp_1})$	$f_{max}(\boldsymbol{xp_1})$	...	$f_{mo}(\boldsymbol{xp_1}, p)$	...	$f_{mo}(\boldsymbol{xp_1}, \alpha)$	$f_{mo}(\boldsymbol{xp_1}, \alpha - 1)$	ch_1
2	$f_\mu(\boldsymbol{xp_2})$	$f_{max}(\boldsymbol{xp_2})$	...	$f_{mo}(\boldsymbol{xp_2}, p)$	...	$f_{mo}(\boldsymbol{xp_2}, \alpha)$	$f_{mo}(\boldsymbol{xp_2}, \alpha - 1)$	ch_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$
m	$f_\mu(\boldsymbol{xp_m})$	$f_{max}(\boldsymbol{xp_m})$	...	$f_{mo}(\boldsymbol{xp_m}, p)$	...	$f_{mo}(\boldsymbol{xp_m}, \alpha)$	$f_{mo}(\boldsymbol{xp_m}, \alpha - 1)$	ch_m

Etapa 4 — Construção de Modelos

Uma vez construída a tabela atributo-valor, busca-se nesta etapa identificar relações entre os atributos, representados por medidas estatísticas e *motifs*, que permitam prever a classe do horizonte de previsão. Uma das maneiras de se identificar essas relações é por meio da construção de modelos de classificação, os quais buscam construir relações matemáticas e lógicas entre os atributos existentes na tabela para identificar a qual classe esses atributos estão associados. Desse modo, o objetivo desta etapa é construir uma hipótese que permita prever as diferentes classes de horizontes.

Nesse sentido, algoritmos de aprendizado de máquina podem dar apoio na indução desses modelos. Aprendizado de máquina consistem em uma subárea de pesquisa em Inteligência Computacional, que tem como objetivo a construção de métodos capazes de adquirir conhecimento de forma automática [Witten and Frank 2005]. É importante ressaltar que, em grande parte das áreas de aplicação, é preferível a construção de modelos simbólicos, como árvores e regras de decisão, de modo que o conhecimento embutido nos modelos possa ser interpretado e compreendido pelos especialistas do domínio.

Etapa 5 — Avaliação de Modelos

Nesta etapa os modelos de classificação construídos na etapa anterior são avaliados e validados por meio de medidas objetivas, juntamente com a participação de especialistas do domínio. No contexto deste trabalho, podem ser destacados dois tipos de erros: de classificação e de previsão. Devido ao fato do problema de previsão ter sido representado como um problema de classificação, primeiramente é analisado o erro de classificação, que pode ser calculado, por exemplo, por meio do método de validação cruzada, que possibilita obter uma aproximação ao erro verdadeiro do modelo [Alpaydin 2004] e pela tabela de contingência, que permite avaliar o relacionamento entre duas ou mais variáveis nominais e, com isso, extrair as medidas de sensibilidade, especificidade, valor preditivo positivo e valor preditivo negativo [Doria 1999]. Em relação ao erro de previsão, devem ser utilizadas técnicas que permitam quantificar a diferença entre os valores observados e os valores previstos utilizando o método. Para isso, são calculados medidas de avaliação como o Erro Médio Absoluto e o coeficiente de correlação. Em [Hyndman and Koehler 2006] é apresentada uma discussão a respeito de diversas medidas utilizadas na literatura para a avaliação de métodos de previsão.

Outra técnica envolvida na avaliação de modelos consiste na avaliação e validação dos modelos por parte do especialista do domínio. A contribuição dos especialistas pode auxiliar na identificação de padrões desnecessários e na certificação dos demais padrões, proporcionando maior robustez preditiva aos modelos construídos. Também podem ser propostas melhorias nas etapas anteriores, de modo a extrair parâmetros mais significativos e, consequentemente, formular hipóteses mais consistentes.

3. Experimentos Preliminares

Para apresentar a ideia do método proposto neste trabalho, foi realizado um experimento preliminar com dados artificiais. Foi construída uma série artificial com medidas estatísticas, *motifs* e comportamentos futuros, conhecidos, de modo a representar claramente o funcionamento do método. A seguir é descrita a construção dessa série e, posteriormente, a aplicação do método proposto de acordo com essa série.

3.1. Construção da Série Artificial

Inicialmente, foi definido um protótipo de 24 dados sequênciais constituído de: (a) nove valores randômicos; (b) cinco valores referentes a um *motif*; (c) cinco valores randômicos; e (d) seis valores referentes ao comportamento futuro. Foram gerados dois *motifs*, mo_1 e mo_2, e quatro classes de comportamentos, c_1, c_2, c_3 e c_4, utilizando sequências de valores randômicos. Os dados randômicos gerados em todo o trabalho são normais com média igual a zero e desvio-padrão igual a um. A partir desse protótipo foram definidas as condições da relação entre os *motifs* e as classes, conforme a Equação 2.

$$\begin{aligned} &\textbf{if } mo_1 \textbf{ and } (f_\mu \geq 0) \Rightarrow c_1 \\ &\textbf{if } mo_2 \textbf{ and } (f_\mu \geq 0) \Rightarrow c_2 \\ &\textbf{if } mo_1 \textbf{ and } (f_\mu < 0) \Rightarrow c_3 \\ &\textbf{if } mo_2 \textbf{ and } (f_\mu < 0) \Rightarrow c_4 \end{aligned} \tag{2}$$

em que f_μ consiste no cálculo da média dos valores das sequências (a), (b) e (c).

Na Figura 1 são apresentados dois gráficos de exemplos de séries construídas utilizando esse protótipo. Nos gráficos é delineado um *motif*, considerando $(f_\mu \geq 0)$ e $(f_\mu < 0)$, bem como as classes correspondentes para cada caso.

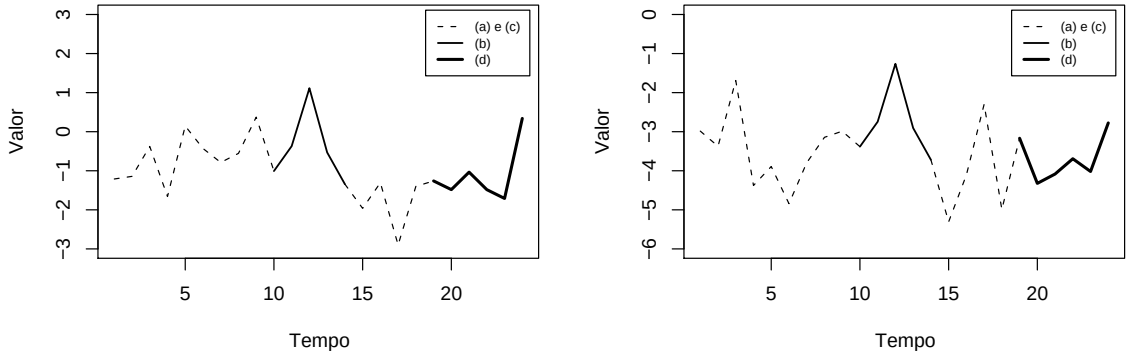

Figura 1. Exemplos de séries construídas de acordo com o protótipo definido.

Desse modo, foram concatenadas 100 séries conforme o protótipo mencionado, alternando aleatoriamente entre os *motifs* e gerada $X = (x_1, x_2, \ldots, x_n)$, em que $n = 2400$. Para a construção da série de dados, bem como a geração dos gráficos deste trabalho, foi utilizado o ambiente computacional R [R Development Core Team 2008]. Na Figura 2, são apresentados os primeiros 300 valores da série artificial gerada.

3.2. Aplicação do Método Proposto

Com base na série artificial gerada foi aplicado o método proposto. Na Etapa (1), definição dos parâmetros iniciais, baseando-se nas informações da série gerada, os parâmetros tamanho de janela (w), horizonte de previsão (h) e passado considerado (p), foram definidos com valores 24, 19 e 6, respectivamente. Com isso, foi construído o conjunto XD contendo os dados estruturados da série X. Após, na Etapa (2), de determinação das classes, devido ao fato das classes serem conhecidas previamente, apenas foi verificada a existência das quatro classes, c_1, c_2, c_3 e c_4. A partir do conjunto de

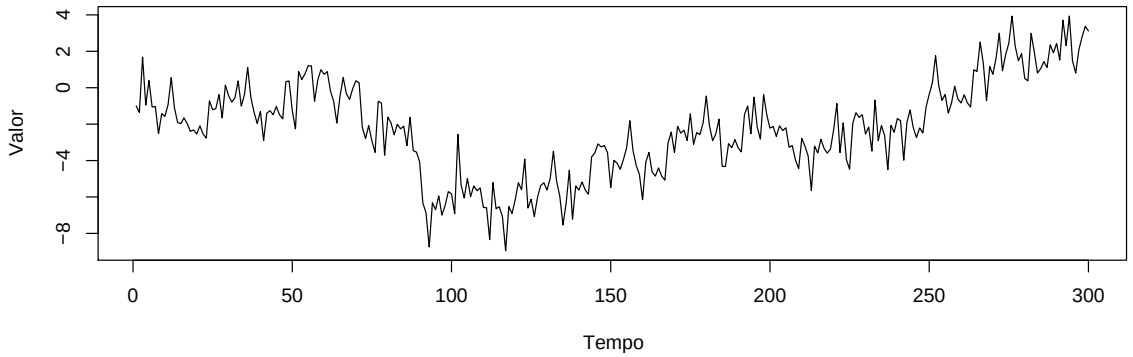

Figura 2. Primeiros 300 valores da série artificial construída.

XD foi construído o conjunto de dados do passado XD', a partir do qual, na Etapa (3), foi realizada a extração de características e de padrões e construída a tabela atributo-valor. Para isso foi utilizada a abordagem proposta em [Maletzke et al. 2009], implementada por meio de um algoritmo desenvolvido na linguagem R [R Development Core Team 2008]. Neste trabalho inicial, foi extraída somente a média como estatística descritiva e procurados os *motifs* de tamanho $\alpha = 5$, que consiste no tamanho de mo_1 e mo_2. Na execução dessa etapa, foram encontrados nove *motifs* ao total, entre os quais, os *motifs* mo_1 e mo_2, inseridos intencionalmente na série temporal.

Na Etapa (4), a tabela atributo-valor foi utilizada para construir um modelo que permitisse representar o relacionamento entre os *motifs* e as classes de comportamentos futuros. Para isso, foi utilizada a ferramenta Weka [Witten and Frank 2005], a qual possibilita a execução de diversos algoritmos de aprendizado de máquina, dentre esses, o algoritmo $J48$, que é uma implementação do algoritmo $C4.5$ proposto em [Quinlan 1993], e consiste na divisão de conjuntos de dados multidimensionais a partir do traçado de sucessivos hiperplanos, que permitam diferenciar as classes envolvidas no problema. Na Figura 3 é apresentada a árvore de decisão resultante da execução do algoritmo $J48$.

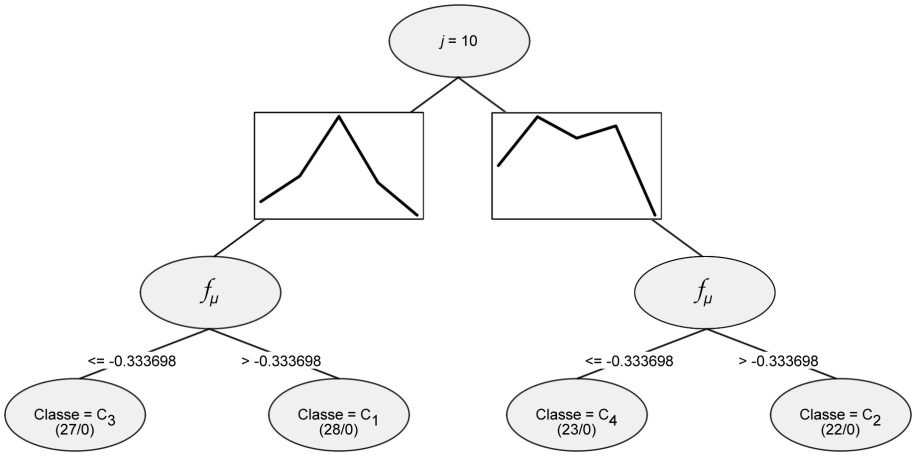

Figura 3. Árvore de decisão induzida pelo algoritmo *J48*.

Com base nessa figura é possível evidenciar os relacionamentos entre os atributos e as classes. De acordo com o esperado, o algoritmo relacionou os *motifs* mo_1 e mo_2 inseridos, juntamente com o resultado da média dos valores do passado, constituídos por (a),

(b) e (c), para a previsão das classes de comportamentos futuros c_1, c_2, c_3 e c_4. Embora tenham sido encontrados outros *motifs*, mo_1 e mo_2 foram os que apresentaram maior representatividade para a indução da árvore de decisão. É importante ressaltar que o traçado dos hiperplanos é realizado por meio do cálculo do valor médio das médias dos exemplos que dividem as classes com maior ganho de informação [Witten and Frank 2005], por esse fato, o valor da média que divide as classes $\{c_1, c_3\}$ e $\{c_2, c_4\}$ não é precisamente zero, mas sim um valor próximo de zero. Neste experimento inicial, não houveram erros de classificação nem de previsão associados ao modelo construído, pois as séries foram geradas artificialmente.

Devido ao fato de que a tabela-atributo valor é construída considerando a divisão da série temporal utilizando a janela w, neste estudo inicial, foi aplicado o método apenas para um instante dessa janela, de modo que seja possível encontrar os relacionamentos conhecidos inseridos na série de dados temporais. Porém, para que possa ser realizada a previsão considerando qualquer instante de tempo, o método deve ser aplicado w vezes para construir modelos referentes a cada possível instante de previsão.

4. Conclusão

Neste trabalho foi apresentado um novo método para previsão de comportamentos temporais, que permite relacionar informações descritivas do passado e padrões (*motifs*) para auxiliar na previsão de comportamentos futuros. Para ilustrar a ideia desse método foi realizado um experimento preliminar utilizando uma série de dados temporais gerada artificialmente. Os relacionamentos inseridos na série de dados artificial foram analisados por meio da utilização de um algoritmo para construção de árvores de decisão. Assim, por meio do método proposto foi possível encontrar e explicar o relacionamento entre as informações do passado e os comportamentos futuros.

Trabalhos futuros incluem a aplicação do método proposto em dados artificiais, utilizando *motifs* e comportamentos futuros de diferentes tamanhos, bem como outros valores para os parâmetros envolvidos; a utilização de outras medidas descritivas que permitam caracterizar comportamentos do passado; a aplicação do método para o estudo de séries de dados temporais de fenômenos reais.

Referências

Alpaydin, E. (2004). *Introduction to Machine Learning*. MIT Press, Cambridge, MA, England.

Antunes, C. M. and Oliveira, A. L. (2001). Temporal data mining: an overview. In *Proceedings of the Workshop on Temporal Data Mining, 7th International Conference on Knowledge Discovery and Data Mining*, pages 1–13, San Francisco, CA, USA.

Buhler, J. and Tompa, M. (2002). Finding motifs using random projections. *Journal of Computational Biology*, 9(2):225–242.

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, NY, USA.

Chun-Hua, B. and Xin-Bao, N. (2004). Determining the minimum embedding dimension of nonlinear time series based on prediction method. *Chinese Physics*, 13(5):633–636.

Doria, U. (1999). *Introdução à Bioestatística: para Simples Mortais*. Elseiver, São Paulo, SP, Brasil.

Everitt, B. S. (1993). *Cluster Analysis*. Edward Arnold, Londres, Inglaterra, 3 edition.

Ferrero, C. A., Monard, M. C., Lee, H. D., and Wu, F. C. (2009). Proposta de uma função de previsão de dados temporais para o algoritmo dos vizinhos mais próximos. In *Anais do XXXV Confêrencia Latinoamericana de Informática (CLEI), A ser publicado*, pages 1–10, Pelotas, RS, Brasil.

Grassberger, P. and Procaccia, I. (1983). Measuring the strangeness of strange attractors. *Physica D: Nonlinear Phenomena*, 9(1–2):189–208.

Hyndman, R. J. and Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679–688.

Karunasinghe, D. S. K. and Liong, S.-Y. (2006). Chaotic time series prediction with a global model: Artificial neural network. *Journal of Hydrology*, 323(1-4):92–105.

Kennel, M. B., Brown, R., and Abarbanel, H. D. I. (1992). Determining embedding dimension for phase-space reconstruction using a geometrical construction. *Physical Review A*, 45:3403–3411.

Kulesh, M., Holschneider, M., and Kurennaya, K. (2008). Adaptive metrics in the nearest neighbours method. *Physica D: Nonlinear Phenomena*, 237(3):283–291.

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.

Maletzke, A. G., Batista, G. E., Lee, H. D., and Wu, F. C. (2009). Mineração de séries temporais por meio da extração de características e da identificação de motifs. In *Anais do VII Encontro Nacional de Inteligência Artificial (ENIA), XXIX Congresso da Sociedade Brasileira de Computação (CSBC)*, pages 637–646, Bento Gonçalves, RS, Brasil.

McNames, J. (1999). *Innovations in Local Modeling for Time Series Prediction*. PhD thesis, Stanford University, Stanford, CA, USA.

Mörchen, F. (2006). Time series knowledge mining. Dissertação de mestrado, Department of Mathematics and Computer Science–Philipps-University, Marburg, Hesse, Germany.

Povinelli, R. J. and Feng, X. (2003). A new temporal pattern identification method for characterization and prediction of complex time series events. *IEEE Transactions on Knowledge and Data Engineering*, 15(2):339–352.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Francisco, CA, USA.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Witten, I. H. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco, CA, USA, 2 edition.