Estudo do Parâmetro Tamanho de *Motif* para a Classificação de Séries Temporais de ECG

André Gustavo Maletzke[1], Huei Diana Lee[1,2], Willian Zalewski[1], Jefferson Tales Oliva[1], Renato Bobsin Machado[1,3], Cláudio Saddy Rodrigues Coy[3], João José Fagundes[3], Feng Chung Wu[1,2,3]

[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
Pós-Graduação em Ciências da Cirurgia
CEP 13083-887 – Campinas, SP, Brasil

{andregustavom,hueidianalee}@gmail.com

Abstract. *Currently, there is a growing interest in different areas for the analysis of data that present temporal dependency. In the medical area, information of this kind is daily recorded, however, only a small portion is analyzed due to lack of methods and tools. In this sense, the identification of morphological patterns (motifs) in time series is an important tool for analyzing such data. In this work, we present an initial study about the influence of the parameter size of the motif considering the classification task, applied to medical temporal databases, such as the electrocardiogram tests. The results show a significant influence of the motif´s size in the classification process.*

Resumo. *Atualmente, existe um crescente interesse em distintas áreas pela análise de dados temporais. Na área médica são registradas diariamente informações com essa característica, no entanto, somente uma pequena parcela é analisada devido à carência de métodos e ferramentas. Nesse sentido, a identificação de padrões morfológicos (motifs) em séries temporais constitui uma ferramenta importante para a análise desses dados. Neste trabalho, é apresentado um estudo inicial sobre a influência do parâmetro tamanho de motif, por meio da tarefa de classificação, quando aplicado em bases de dados temporais da área médica, como exames de Eletrocardiograma. Observou-se influência significativa desse parâmetro em relação aos resultados de classificação.*

1. Introdução

O desenvolvimento de tecnologias computacionais ao longo dos anos tem aumentado de maneira acelerada a capacidade de adquirir e armazenar informações [Hilbert and López 2011]. No entanto, o poder de análise de informação não tem seguido o mesmo ritmo de crescimento. Esse fato pode ser observado em distintas áreas

como na medicina, na qual, diariamente, inúmeros equipamentos médicos capturam informações sobre variáveis físicas, químicas e biológicas que descrevem o quadro clínico de pacientes. Porém, somente uma pequena parcela dessas informações é armazenada para posterior análise de um determinado momento ou ao longo do tempo, seja de modo manual ou por meio de ferramentas computacionais.

Um exemplo são os exames de Eletrocardiograma – ECG –, os quais consistem na monitoração ao longo do tempo da variação dos potenciais elétricos gerados pela atividade cardíaca e são de fundamental importância para o diagnóstico de um grande número de cardiopatias e outros distúrbios [Thanapatay et al. 2010].

De acordo com a Organização Mundial da Saúde[1] as doenças cardiovasculares são a principal causa de morte no mundo, representando cerca 29% dos casos. Segundo o Ministério da Saúde, as doenças cardiovasculares também constituem uma das principais causas de morte no Brasil, totalizando 32% dos óbtidos bem definidos, sendo que esses índices atingem estratos populacionais mais jovens de maneira mais acentuada quando comparados a outros países como Estados Unidos, Japão e Europa Ocidental. Apenas em 2005, 22% dos gastos com internação hospitalar (exceto partos) estão ligados a doenças cardiovasculares seguido por doenças respiratórias crônicas (15%) e neoplasias (11%) [Ministério da Saúde 2009], estatísticas que ressaltam a necessidade de políticas públicas mais eficientes para a prevenção e o diagnóstico dessa enfermidade.

Nesse contexto, métodos computacionais podem prover suporte à análise de dados obtidos ao longo do tempo como exames de ECG podendo auxiliar na identificação de padrões importantes de modo a contribuir na realização de diagnósticos precoces e com maior precisão. Todavia, grande parte dos métodos computacionais disponíveis aos especialistas são destinados à análise de dados não temporais. Desse modo, é necessário que técnicas específicas sejam desenvolvidas e aplicadas a dados com essa característica.

Dentre as técnicas existentes, a identificação de padrões morfológicos, denominados de *motifs*, em Séries Temporais – ST – tem sido utilizada em distintas áreas por meio do apoio a tarefas como descrição de dados, classificação, predição, entre outras [Mitchell 1997, Monard and Baranauskas 2003]. Assim sendo, a identificação de *motifs* em exames com coleta de dados ao longo do tempo constitui uma ferramenta importante podendo contribuir significativamente no estudo desses dados. No entanto, a definição do parâmetro tamanho de *motif* é um processo complexo devido ao grande número de possibilidades que podem ser consideradas para cada problema e está diretamente relacionada ao fracasso ou ao sucesso da análise que se pretende realizar. Portanto, o estudo e o entendimento desse parâmetro é uma tarefa de interesse e necessária para a utilização de *motifs* em problemas reais.

A identificação de *motifs* é uma tarefa custosa computacionalmente, podendo se tornar inviável em situações nas quais as ST possuem número elevado de observações, pois o método tradicional apresenta complexidade quadrática em relação ao número de observações da ST [Maletzke and Batista 2010]. Neste trabalho, é utilizado um método probabilístico baseado em [Chiu et al. 2003] e adaptado em [Maletzke 2009] para a construção de tabelas atributo-valor para posterior aplicação de métodos de extração de conhecimento. Embora apresente um comportamento probabilístico, em [Maletzke et al. 2008]

[1] `http://www.who.int`

é apresentada uma avaliação experimental ilustrando a eficiência do método.

Este trabalho faz parte do projeto de Análise Inteligente de Dados desenvolvido mediante uma parceria entre o Laboratório de Bioinformática – LABI – da Universidade Estadual do Oeste do Paraná – UNIOESTE/Foz do Iguaçu, o Serviço de Coloproctologia da Faculdade de Ciências Médicas da Universidade Estadual de Campinas – UNICAMP/Campinas, o Laboratório de Inteligência Computacional – LABIC – do Instituto de Ciências Matemáticas e de Computação – ICMC – da Universidade de São Paulo – USP/São Carlos e o Grupo Interdisciplinar em Mineração de Dados e Aplicações – GIMDA – da Universidade Federal do ABC – UFABC, e tem como objetivo realizar um estudo inicial referente à influência do parâmetro tamanho de *motif* aplicado a base de dados temporais de exames médicos, com um estudo de caso de ECG.

O restante deste trabalho está organizado da seguinte maneira: nas Seções 2 e 3 são apresentados os trabalhos relacionados e o material e método, respectivamente. Na Seção 4 são apresentados e discutidos os resultados e na Seção 5 são apresentadas as conclusões e os trabalhos futuros.

2. Trabalhos Relacionados

O desenvolvimento de métodos para auxiliar na análise de dados oriundos de exames de ECG tem sido tema de estudo de inúmeros trabalhos, sendo que várias abordagens foram propostas envolvendo problemas como predição, descrição/sumarização, classificação e visualização de dados de ECG. Em relação ao problema de classificação grande parte dos trabalhos destina-se à extração de características relevantes para posterior indução de classificadores baseados, principalmente, em redes neuronais artificiais [Neagoe et al. 2003, Thanapatay et al. 2010, Mar et al. 2011].

Em [Yu and Chen 2007] os exames de ECG são representados em função dos coeficientes obtidos a partir da transformada discreta de *wavelet* e utilizados como dados de entrada para construção de uma rede neuronal artificial. Os resultados são promissores, no entanto, a inteligibilidade dos modelos construídos é bastante complexa e pouco intuitiva, prejudicando a utilização do conhecimento por parte dos especialistas.

Em [Jovic and Bogunovic 2010] é realizada a extração de características baseadas em medidas estatísticas e em medidas como dimensão de correlação e entropia para indução de distintos classificadores dentre os quais redes neuronais artificiais, árvores de decisão e máquinas de vetor e suporte. A indução de árvores de decisão apresentou resultados significativos e promissores, permitindo que o conhecimento gerado fosse analisado por meio das regras geradas. Embora os resultados tenham sido promissores as características utilizadas apresentam baixa inteligibilidade.

A abordagem de *ensemble* para combinar vários classificadores neuronais em um sistema eficiente para a classificação de exames de ECG é proposta em [Osowski et al. 2011], sendo que a combinação dos classificadores é realizada por meio da utilização de algoritmos genéticos. Os resultados da aplicação dessa abordagem foram superiores quando comparados à abordagem sem *ensemble*.

De modo geral, a maioria dos trabalhos combinam extração de características e diferentes algoritmos de aprendizado de máquina para aumentar as taxas de classificação de problemas envolvendo dados de ECG. Nesse sentido, muitos trabalhos tem alcançados

resultados importantes e relevantes para a área, no entanto, pouco esforço tem sido destinado à construção de classificadores inteligíveis permitindo e apoiando a especialistas no estudo e entendimento de distúrbios cardiovasculares.

3. Material e Método

Nesta seção são descritos o método utilizado para a identificação de padrões morfológicos, o conjunto de dados e o método aplicado para avaliar o parâmetro tamanho de *motif*.

3.1. Descrição do Conjunto de Dados

O Eletrocardiograma é um exame que tem por objetivo descrever ao longo do tempo os fenômenos elétricos relacionados à atividade cardíaca por meio de eletrodos pré-posicionados no corpo. Esse exame quando realizado de modo completo é constituído por doze eletrodos, no entanto, aparelhos que realizam a análise de maneira automática, frequentemente utilizam somente uma parte desses eletrodos [Olszewski 2001]. O conjunto de dados selecionado foi obtido do repositório de dados da UCR *Time Series Classification/Clustering*[2] e está composto por 200 exames de ECG de diferentes pacientes. Na Tabela 1 é apresentada uma breve descrição a respeito do conjunto de dados, na qual as colunas 2 e 3 representam o quantidade de exemplos, *i.e.*, a número de exames, e o tamanho de cada exemplo, respectivamente. As colunas 4 e 5 referem-se ao número e distribuição de classes e as colunas 6 e 7 ao erro da classe majoritária e a indicação da existência ou não de valores desconhecidos nos dados.

Tabela 1. Características do conjunto de dados de ECG

Conjunto de dados	# Ex.	#Observações da ST	Classes	% Classe	Erro majoritário	Valores desconhecidos
ECG	200	96	1 2	66,5% 33,5%	33,5%	Nenhum

Os exames de ECG que constituem o conjunto de dados referem-se a pacientes que apresentaram sinal clínico de taquicardia supra-ventricular (classe 2) e não supra-ventricular (classe 1), previamente analisados por especialistas do domínio. No gráfico da Figura 1(a) é apresentado um exame com sinal clínico de taquicardia não supra-ventricular e no gráfico da Figura 1(b)com sinal clínico de taquicardia supra-ventricular.

3.2. Método para a Identificação de Padrões Morfológicos (*motifs*)

Como mencionado, a identificação de *motifs* é uma tarefa custosa computacionalmente podendo apresentar, no pior caso, uma complexidade de $O(m^2)$, em que m é o tamanho da ST. Desse modo, buscando-se reduzir a complexidade computacional foi utilizado um método baseado na abordagem desenvolvida em [Chiu et al. 2003] e adaptada em [Maletzke 2009] para aplicação em ST envolvendo problemas de classificação. O método utilizado pode ser divido em três passos, apresentados a seguir.

3.2.1. 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. Esses conceitos são apresentados na Definição 1 e Definição 2, respectivamente.

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

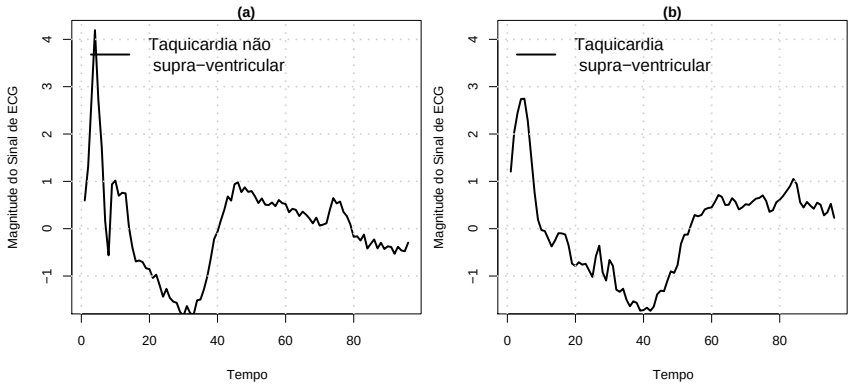

Figura 1. Exemplos de exames de ECG extraídos do conjunto de dados

Definição 1 *(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 $1 \leq p \leq m - n + 1$.*

Definição 2 *(Janela Deslizante) [Maletzke 2009] 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 $1 \leq i \leq m-n+1$.*

Desse modo, uma vez construída a MS é aplicado o algoritmo *Random Projections* [Buhler and Tompa 2002] com o objetivo de identificar subsequências similares. Porém, esse algoritmo foi originalmente proposto para sequências de DNA e de proteínas e requer como entrada uma sequência de símbolos. Para tanto, no método utilizado, cada subsequência da MS é transformada para uma subsequência simbólica (cada ponto da subsequência correspondendo a um valor do alfabeto), de acordo com a Definição 3, por meio do método *Symbolic Aggregate aproXimation* – SAX – [Lin et al. 2002].

Definição 3 *(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 $\hat{z}_t \in \Sigma$, onde Σ é um alfabeto finito.*

Uma variação que pode ser realizada, é utilizar o SAX, em uma etapa preliminar, para a redução de dimensionalidade por meio da definição de uma janela de redução, a qual pode ser ajustada em relação ao tamanho da sequência. Após, a sequência é discretizada utilizando um determinado alfabeto. No exemplo da Figura 2 (a) são extraídas subsequências de tamanho 16 da ST Z, com tamanho $m = 1000$, que após realizada a redução de dimensionalidade por meio de uma janela de redução de 25% e discretizadas mediante o alfabeto (a,b,c) formam subsequências simbólicas de tamanho $n_{sax} = 4$.

3.2.2. Passo 2 – Construção da Matriz de Colisão

A Matriz de Colisão – MC – é utilizada como ferramenta 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

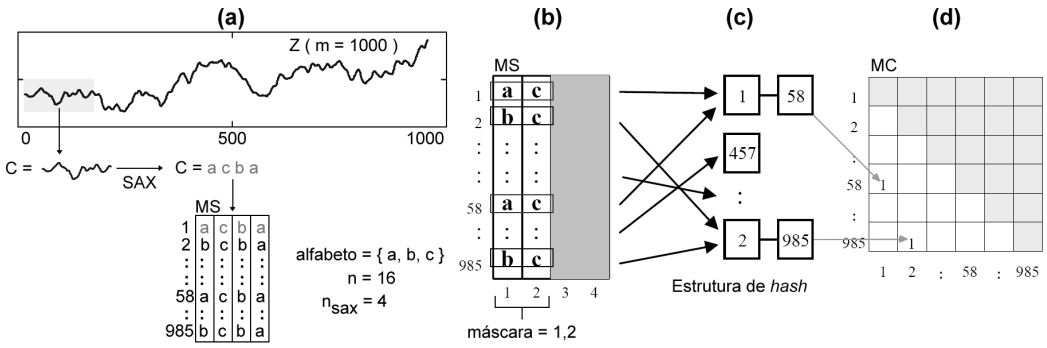

Figura 2. Processo de identificação de *motifs* (modificada de [Chiu et al. 2003])

é igual a $(1{,}2)$, portanto as subsequências das linhas $(1{,}58)$ e $(2, 985)$ da MS irão colidir, pois ambas possuem os mesmos valores nas colunas 1 e 2. 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, com o objetivo de tornar as máscaras mais representativas.

3.2.3. Passo 3 – Análise da Matriz de Colisão

Um valor alto em uma posição da MC é um indício da existência de um *motif*. Para que esse indício apontado pela MC seja comprovado 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, mais especificamente neste trabalho a distância Euclidiana. Caso duas subsequências estejam dentro de uma distância r, essas podem ser consideradas como *motifs*. Neste método, o limiar r refere-se a uma percentagem que corresponde ao erro médio aceito para a diferença existente entre duas observações de duas subsequências distintas.

Para que seja determinado um valor em percentagem condizente com o valor de r é necessário que as subsequências sejam normalizadas para o intervalo 0 e 1 para posterior determinação da distância. Esse procedimento permite que seja desconsiderada a informação de amplitude dos dados em relação aos valores que compõem as sequências comparadas, apresentando resultados eficientes na comparação da morfologia. Desse modo, uma vez comprovada a similaridade morfológica é necessário verificar se as subsequências apresentam amplitudes de valores similares. Para isso, são resgatados os valores, sem normalização, das subsequências e determinada a área de cada subsequência, de modo que subsequências com áreas que diferem em no máximo r serão consideradas similares. O cálculo da área é realizado por meio da regra dos trapézios [Barroso et al. 1987].

Outras subsequências também podem estar dentro de r, porém podem não ter sido selecionadas devido à característica probabilística do método e, portanto, precisam ser adicionadas à condição de *motif*. Neste trabalho, é realizada uma busca sequencial a partir da subsequência definida como *motif* por toda a ST.

3.3. Avaliação Experimental

Para avaliar o parâmetro tamanho de *motif* envolvendo dados de exames de ECG foram extraídos *motifs* de distintos tamanhos mediante o método apresentado na Seção 3.2. A

definição do tamanho de *motifs* a serem extraídos foi realizada considerando uma percentagem em relação ao número total de observações de cada exame, o qual é equivalente a 96 observações. Desse modo, para realizar o processo de identificação de *motifs* foi utilizada a seguinte configuração experimental:

- **Tamanho de *motif*:** foram extraídos *motifs* nos intervalos de 5% a 50%, com incrementos de 5%, em relação ao número total de observações dos exames;
- **Limiar de similaridade:** foi utilizado um limiar r de 5%, isto é, as subsequências selecionadas como *motifs* diferem em no máximo 5%, considerando conjuntamente a combinação entre a distância Euclidiana e a área de cada subsequência;
- **Janela de redução de dimensionalidade:** não foi aplicado nenhum método de redução de dimensionalidade;
- **Tamanho da máscara:** utilizou-se uma máscara de tamanho dois corroborando com estudos realizados em [Chiu et al. 2003] com distintos conjuntos de dados;
- **Tamanho do alfabeto:** o alfabeto utilizado é composto por seis símbolos;
- **Número de iterações:** as iterações realizadas corresponderam a 50% de todas as possíveis combinações de máscaras.

Para cada variação do parâmetro tamanho de *motif* foi construída uma tabela atributo-valor, na qual cada *motif* de um mesmo tamanho, porém de morfologias distintas, constitui um atributo cujo valor é dado pela frequência desse *motif* em cada exame. Portanto, considerando o intervalo selecionado para se realizar a variação desse parâmetro foram obtidas 10 tabelas atributo-valor.

Para determinar a capacidade de descrição das tabelas atributo-valor obtidas da variação do parâmetro tamanho de *motif* foram realizadas a indução de classificadores, um para cada tabela atributo-valor, e mensurado o erro de classificação de cada classificador. Para a construção dos classificadores foi utilizado o algoritmo *J*48, por meio da ferramenta WEKA[3], para indução de árvores de decisão.

Para reduzir a possibilidade de que os resultados tenham sido gerados ao acaso foi utilizado o método de amostragem denominado de validação cruzada 5×2. Esse método consiste em particionar cinco vezes o conjunto de dados, considerando sementes diferentes, gerando um conjunto de teste e outro de treinamento a cada vez, de modo a preservarem as características do conjunto de dados original.

A análise dos resultados foi realizada por meio do teste estatístico ANOVA com pós-teste Tukey para comparações múltiplas [Motulsky 1995].

4. Resultados e Discussão

Na Tabela 2 são apresentadas as taxas médias de erro e os respectivos desvios padrão dos classificadores gerados pelo algoritmo *J*48 para cada variação do tamanho de *motif*.

Na análise dos resultados não foi observada diferença estatisticamente significativa entre os erros de classificação obtidos para cada variação de tamanho de *motif* (*p-valor* $> 0{,}05$), exceto para o tamanho de *motif* 5% que apresentou diferença estatisticamente significativa em relação aos demais tamanhos (*p-valor* $< 0{,}001$). Observou-se que o tamanho do *motif* influencia no desempenho do modelo e que, embora não tenha sido

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

possível detectar diferença estatísticamente significativa entre os tamanhos 10% a 50%, houve uma têndência de maiores erros para os tamanhos 45% e 50%. Esse fato evidencia a importância de se investigar a existência de comportamentos locais nos dados.

Tabela 2. Taxas de erro médio e desvios padrão para cada tamanho de *motif*

	Tamanhos de *Motif* em Percentagem									
	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%
Erro Médio	9,61	25,08	27,51	25,34	24,82	23,09	25,09	25,70	28,00	28,31
DP	3,66	5,79	5,36	6,35	5,42	3,06	3,79	2,24	4,12	2,34

Muitas das análises em exames de ECG são realizadas de maneira global por meio, por exemplo, de características que buscam descrever cada exame. A análise de pequenos comportamentos pode auxiliar na identificação de anomalias ainda em estágios iniciais. Portanto, neste estudo os resultados apontaram uma maior relevância, considerando a classificação de exames de ECG, para padrões morfológicos de menor tamanho em relação ao número de observações dos exames. Nesse sentido, acredita-se que os resultados ainda podem ser melhorados por meio da busca de *motifs* de outros tamanhos, porém restritos a tamanhos menores corroborando com esses resultados iniciais apresentados.

Em relação ao método utilizado para a identificação de *motifs*, este apresenta melhor eficiência computacional quando comparado ao método denominado de força bruta [Lin et al. 2002], no entanto possui uma quantidade maior de parâmetros. Dentre esses parâmetros, o tamanho de *motif*, a medida de similaridade e o limiar de similaridade são considerados de maior importância, pois estão diretamente relacionados à aceitação de subsequências como sendo similares ou não. Para realizar a avaliação do parâmetro tamanho *motif* variou-se o tamanho desse parâmetro de maneira crescente buscando identificar padrões locais nos dados e não uma morfologia global que os descreva.

Já para o parâmetro limiar de similaridade optou-se por abrandar o nível de erro de 0% para 5% considerando que com um limiar de 0% não foi possível identificar padrões morfológicos nos dados que pudessem dar apoio à tarefa de classificação. Esse parâmetro é um dos mais complexos de ser determinado, pois está diretamente relacionado com a geração de falsos positivos e falsos negativos, sendo necessário um estudo mais detalhado a respeito em trabalhos futuros. Juntamente com esse parâmetro a medida de similaridade é outro parâmetro a ser ajustado, embora estudos realizados na literatura apontem a distância Euclidiana como sendo em média a que apresenta resultados melhores [Keogh and Kasetty 2002], existem outras métricas ou combinações de métricas que podem ser utilizadas para a comparação de subsequências.

Os demais parâmetros são utilizados para reduzir o espaço de busca e não possuem influência na aceitação de subsequências como *motifs*, *i.e.*, não estão diretamente relacionados à geração de falsos positivos. Desse modo, o número de iterações foi fixado em 50% do total de iterações possíveis, considerando que em estudos anteriores foi verificada a eficiência do método com valores menores [Maletzke 2009]. O tamanho do alfabeto foi determinado considerando estudos anteriores [Lin et al. 2002].

Em relação à amostragem de dados, foi escolhida a validação cruzada 5×2, pois o conjunto de dados possui poucos exemplos e apresenta classes desbalanceadas, podendo o resultado ser comprometido se utilizada a validação cruzada com dez partições.

É importante ressaltar que os resultados apresentados podem variar de acordo com

o algoritmo de indução de classificadores. Neste trabalho, foi realizada a indução de árvores de decisão considerando que o conhecimento gerado por algoritmos dessa classe apresentam maior inteligibilidade em relação a algoritmos de outros paradigmas, permitindo com que o raciocínio utilizado na classificação seja facilmente explicado em função dos *motifs* encontrados nos dados do exame [Mitchell 1997, Monard and Baranauskas 2003].

5. Conclusão e Trabalhos Futuros

De acordo com os resultados apresentados foi possível observar que a aplicação de *motifs* para a classificação de dados de ECG é promissora apresentando resultados interessantes em termos de taxa de erro e no nível de inteligibilidade dos modelos gerados. Neste trabalho, observou-se que *motifs* de menores tamanhos possuem maior capacidade de descrição quando utilizados em problemas de classificação de dados de ECG. Além disso, a definição do tamanho de *motif* não é uma tarefa trivial sendo necessário realizar análises mais completas para auxiliar na definição desse parâmetro.

Desse modo, espera-se que o estudo de padrões morfológicos locais em exames de ECG possa, futuramente, auxiliar a médicos e especialistas, bem como a centros de atendimento hospitalar, que não contam com especialistas da área, na identificação anormalidades com maior precisão e antecedência.

Como trabalhos futuros pretende-se explorar métodos para a identificação automática de *motifs*, bem como a avaliação de outros parâmetros relacionados ao processo de identificação de padrões morfológicos. Outro aspecto a ser estudado refere-se à construção de tabelas atributo-valor contendo *motifs* de tamanhos distintos de modo que os modelos construídos possam relacionar desde comportamentos locais até comportamentos mais gerais e também estudar o potencial existente na identificação de *motifs* quando aplicado a outros métodos de classificação. Ainda, um outro trabalho futuro será a aplicação dessa abordagem a dados médicos de outras especialidades, como dados provindos de exames de manometria anorretal.

Agradecimentos

A Fundação Araucária e a Universidade Estadual do Oeste do Paraná pelo auxílio na realização deste trabalho por meio do programa de bolsas de iniciação científica.

Referências

Barroso, L. C., de Araújo, M. M., Filho, F. F., de Carvalho, M. L. B., and Maia, M. L. (1987). *Cálculo Numérico*. Harbra.

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, USA.

Hilbert, M. and López, P. (2011). The world's technological capacity to store, communicate, and compute information. *Sciene Magazine*.

Jovic, A. and Bogunovic, N. (2010). Electrocardiogram analysis using a combination of statistical, geometric, and nonlinear heart rate variability features. *Artificial Intelligence in Medicine*, In Press, Corrected Proof.

Keogh, E. and Kasetty, S. (2002). On the need for time series data mining benchmarks: a survey and empirical demonstration. In *Proceedings of the 8th International Conference on Knowledge Discovery and Data Mining*, pages 102–110, New York, USA.

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, Canada.

Maletzke, A. G. (2009). Uma metodologia para extração de conhecimento em séries temporais por meio da identificação de motifs e da extração de características. Dissertação de mestrado, Universidade de São Paulo, São Carlos, São Paulo, Brasil.

Maletzke, A. G. and Batista, G. E. (2010). Mineração de dados temporais mediante a identificação de motifs e a extração de características. In *Anais do VII Best MSc Dissertation/PhD Thesis Contest - Joint Conference 2010*, volume 1, pages 1–12, São Bernardo do Campo - SP.

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

Mar, T., Zaunseder, S., Cortes, J. P. M., Soria, M. L., and Poll, R. (2011). Optimization of ecg classification by means of feature selection. *Biomedical Engineering*, (99):1.

Ministério da Saúde (2009). Elsa brasil: maior estudo epidemiológico da américa latina. *Revista Saúde Pública*, 43(1).

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill, Boston, USA.

Monard, M. C. and Baranauskas, J. A. (2003). *Sistemas Inteligentes: fundamentos e aplicações*, chapter Conceitos sobre Aprendizado de Máquina, pages 89–114. Editora Manole, Barueri, Brasil.

Motulsky, H. (1995). *Intuitive Biostatistics*. Oxford University Press, New York, USA.

Neagoe, V.-E., Iatan, I.-F., and Grunwald, S. (2003). A neuro-fuzzy approach to classification of ecg signals for ischemic heart disease diagnosis. In *Proceedings of the Annual Symposium Proceedings Archive at the Americam Medical Informatics Association*, pages 494–498.

Olszewski, R. T. (2001). *Generalized Feature Extraction for Structural Pattern Recognition in Time-Series Data*. PhD thesis, Carnegie Mellon University, Pittsburgh, PA.

Osowski, S., Siwek, K., and Siroic, R. (2011). Neural system for heartbeats recognition using genetically integrated ensemble of classifiers. *Computers in Biology and Medicine*, 41(3):173 – 180.

Thanapatay, D., Suwansaroj, C., and Thanawattano, C. (2010). Ecg beat classification method for ecg printout with principle components analysis and support vector machines. In *Proceedings of The International Conference on Electronics and Information Engineering*, pages 72–75.

Yu, S.-N. and Chen, Y.-H. (2007). Electrocardiogram beat classification based on wavelet transformation and probabilistic neural network. *Pattern Recognition Letters*, 28:1142–1150.