IDENTIFICAÇÃO DE MOTIFS PARA A CLASSIFICAÇÃO DE SÉRIES TEMPORAIS DE EXAMES DE ECG

Jefferson Tales Oliva (PIBIC/Fundação Araucária/UNIOESTE),
André Gustavo Maletzke, Huei Diana Lee, Feng Chung Wu,
Carlos Andrés Ferrero (Orientador), e-mail: anfer86@gmail.com

Universidade Estadual do Oeste do Paraná/ Centro de Engenharias e Ciências Exatas/ Campus de Foz do Iguaçu/ PR

Palavras-chave: aprendizado de máquina, padrões morfológicos, exames de eletrocardiograma.

Ciências Exatas e da Terra - Ciência da Computação

Resumo

O desenvolvimento tecnológico atrelado ao aumento na capacidade de armazenamento e processamento tem permitido, cada vez mais, que a análise de dados seja apoiada por métodos computacionais. Nesse sentido, o estudo de fenômenos temporais, em séries temporais, tem despertando interesse de distintas áreas, inclusive a área de medicina. Um dos métodos utilizados consiste na extração de padrões morfológicos existentes nos dados. Neste trabalho, é apresentado um método para a extração de *motifs* aplicado a exames de Eletrocardiograma e avaliado mediante a tarefa de classificação.

Introdução

Avanços tecnológicos ao longo dos anos na área da computação impulsionaram o aumento da capacidade de processamento e armazenamento de dados, incentivando o desenvolvimento de sistemas que auxiliem na coleta e na organização de dados em diversos domínios.

Nesse contexto, hospitais e clínicas médicas registram informações sobre pacientes diariamente, sendo que parte dessas informações são geradas por equipamentos médicos. Um dos métodos comumente utilizado é o exame de Eletrocardiograma (ECG), que tem como finalidade monitorar os batimentos cardíacos do paciente, possibilitando identificar uma ampla variedade de doenças e distúrbios cardíacos por meio da análise das variações dos potenciais elétricos gerados ao longo do tempo. Os dados gerados pelo ECG podem ser representados por meio de Séries Temporais (ST), possibilitando que inúmeros métodos de análise dados sejam aplicados. Um dos métodos consiste na identificação de padrões morfológicos, denominados *motifs*, os quais podem ser utilizados como atributos descritivos de uma ST. No entanto, devido à grande quantidade de

dados gerados diariamente a análise desses padrões também se torna uma tarefa consideravelmente complexa. Desse modo, métodos computacionais como o processo de Mineração de Dados, apoiado por técnicas de Aprendizado Máquina (AM), podem ser aplicados para dar suporte na análise desses dados.

Nesse sentido, o objetivo deste trabalho consiste em avaliar o processo de identificação *motifs* mediante distintos métodos de AM quando aplicados a dados de exames de ECG. Este trabalho faz parte do projeto de Análise Inteligente de Dados desenvolvido mediante a parceria entre o Laboratório de Bioinformática (LABI/UNIOESTE), Serviço de Coloproctologia da Faculdade Ciências Médicas da UNICAMP, Laboratório de Inteligência Computacional (LABIC/USP) e o Grupo Interdisciplinar em Mineração de Dados e Aplicações (GIMDA/UFABC).

Materiais e Métodos

O método utilizado para a realização deste trabalho é composto por quatro etapas: (1) determinação do conjunto de dados, (2) identificação de *motifs*, (3) construção de modelos e (4) avaliação dos resultados.

Na Etapa (1), foi selecionado conjunto de dados de séries temporais provenientes do repositório de dados da UCR *Time Series Classification/Clustering* [1], o qual é de domínio público. Esse conjunto de dados é composto por 200 exames de ECG de diferentes pacientes, em que cada exame é composto por uma série temporal com 96 observações e não possuem valores ausentes. De acordo com [2], esses exames foram previamente analisados e classificados como normal (65,3% dos casos) e anormal (33,5% dos casos) por especialistas do domínio. Para a realização da Etapa (2), foi utilizado um método baseado no apresentado em [3], o qual é composto pelas seguintes passos, considerando uma única ST:

- **Passo 1 – Construção da matriz de subsequências:** é realizada a extração de todas as subsequências de um determinado tamanho da ST, por meio do conceito de janela deslizante. Após, cada subsequência é transformada em cadeia de símbolos por meio do método *Symbolic Aggregate aproXimation* (SAX) [4];
- **Passo 2 – Construção da matriz de colisão:** a partir da matriz de subsequências simbólicas e por meio de um processo iterativo são identificadas as subsequências simbólicas similares mediante a aplicação de uma função de *hashing*. Para cada colisão gerada pela função de *hashing*, um contador nas posições correspondentes às subsequências que colidiram é incrementado;
- **Passo 3 – Análise da matriz de colisão:** as posições da matriz de colisão que obtiveram maior número de colisões indicam possíveis *motifs*, os quais são verificados mediante uma medida de similaridade de um limiar de similaridade previamente definido. Assim, as subsequências

que obtiveram valores menores ou iguais a esse limiar, são consideradas *motifs*.

Neste trabalho, foi utilizada a seguinte configuração experimental:

- Tamanho de *motif*: foram extraídos *motifs* de tamanhos 3, 5, 7, 10 e 12 corroborando com os melhores resultados apresentados em [3];
- Medida de similaridade: distância Euclidiana;
- Limiar de similaridade: utilizou-se um limiar de 5%, ou seja, as subsequências selecionadas como *motifs* diferem em no máximo 5%;
- Tamanho do alfabeto utilizado pelo SAX: optou-se por um alfabeto de seis símbolos;
- Número de iterações: equivalente a 50% da quantidade máxima.

Desse modo, após a aplicação da Etapa (2), as ST referentes aos exames de ECG passam a ser representadas por uma tabela atributo-valor, em função da frequência dos *motifs* identificados, em distintos tamanhos.

Na Etapa (3), a partir da tabela atributo-valor construída na etapa anterior, foram construídos modelos de classificação por meio de distintas técnicas de AM. As técnicas utilizadas foram Árvore de Decisão (AD), Vizinhos mais Próximos (VMP) e Redes Neuronais Artificiais (RNA). Para a construção dos classificadores, foi utilizada a ferramenta WEKA[4].

Para reduzir a possibilidade dos resultados serem gerados ao acaso, o conjunto de dados foi dividido em duas partições, uma de treino e outra de teste, para a indução dos classificadores. Esse processo foi repetido cinco vezes. Para a implementação do método de identificação de *motifs*, bem como a realização dos experimentos, foi utilizado o ambiente matemático e estatístico R.

Para a avaliação dos resultados na Etapa (4), foi considerado o erro médio de classificação e a análise dos resultados foi realizada por meio do teste estatístico ANOVA.

Resultados e Discussão

O método foi aplicado a um conjunto de 200 exames de ECG de diferentes pacientes. Desse modo, para a identificação de *motifs*, foram definidas varáveis como o tamanho, a medida de similaridade, o limiar de similaridade, o tamanho do alfabeto utilizado pelo SAX e o número de iterações. Após a aplicação do método de identificação de *motifs* obteve-se uma tabela atributo-valor cujos atributos representam os *motifs* encontrados. Para a construção de modelos, foi utilizada a ferramenta WEKA, que fornece vários algoritmos de AM para a indução de classificadores, tais como: J48 para construção de AD; 1NN para classificação por vizinhos mais próximos; e MLP para construção de RNA.

Os valores de erro médio e desvio-padrão para a indução de AD foram de *0,3069* e *0,4349*; para VMP foram de *0,3306* e *0,0216*; e para RNA foram de *0,2970* e *0,0363*. Para verificar a existência de diferença estatisticamente significativa, foi aplicado o teste estatístico ANOVA, considerando o nível de significância de *95%*. O *p-valor* resultante foi de *0,3299*, que não constatou diferença estatisticamente significativa.

Conclusões

De acordo com os resultados as três técnicas utilizadas não apresentaram diferença estatisticamente significativa quanto ao desempenho na classificação de dados de EGC. Sendo assim, trabalhos futuros incluem a avaliação dos parâmetros do processo de identificação de padrões morfológicos e a aplicação dessa abordagem a séries temporais da área médica em outras especialidades, como exames de manometria anorretal.

Agradecimentos

À Fundação Araucária pelo apoio por meio da linha de concessão de bolsas de iniciação científica, categoria PIBIC/Fundação Araucária/Unioeste.

Referências

1. KEOGH, E.; XI, X.; WEI, L.; RATANAMAHATANA, C.A. **The UCR Time Series Classification/Clustering**. Disponível em: www.cs.ucr.edu/~eamonn/time_series_data/. Acesso em: 23 de Abril de 2010.
2. MALETZKE, A.G.; LEE, H.D.; ZALEWSKI, W.; OLIVA, J.T.; MACHADO, R.B.; COY, C.S.R.; FAGUNDES, J.J.; WU, F.C. **Estudo do Parâmetro Tamanho do Motif para a Classificação de Séries Temporais de ECG**. XI Workshop de Informática Médica – Natal – Rio Grande do Norte. Anais do Workshop de Informática Médica, Natal: UFRN, 2011.
3. MALETZKE, A.G. **Uma metodologia para extração de conhecimento em séries temporais por meio da identificação de motifs e da extração de características**. São Carlos, (Dissertação de mestrado – Universidade de São Paulo – ICMC-USP), 2009.
4. WITTEN, I.H.; FRANK, E. **Data Mining: Practical Machine Learning Tools and Techniques**. 2 Edição. San Francisco: Elsevier, 2005. 525 p.