22º EAIC Encontro Anual de Iniciação Científica
3º EAITI Encontro Anual de Iniciação Tecnológica e Inovação

ESTUDO DO ALGORITMO KNN-TSP PARA PREVISÃO DE SÉRIES TEMPORAIS ARTIFICIAIS E REAIS

Leandro Borges dos Santos(PIBIC/Ações Afirmativas/Fundação Araucária/Unioeste), André Gustavo Maletzke(Orientador), Huei Diana Lee, Richardson Floriani Voltolini, Joylan Nunes Maciel, Renato Bobsin Machado, Wu Feng Chung, e-mail: sborgesleandro@gmail.com

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

Ciências Exatas e da Terra – Ciência da Computação

Palavras-chave: Análise de Dados, Aprendizado de Máquina, Medidas de Similaridade

Resumo: Uma das abordagens para a previsão de dados temporais é utilização de conceitos de Aprendizado de Máquina. Nesse sentido, neste trabalho foi estudo um desses métodos, o kNN-TSP, para avaliar o seu desempenho, este algoritmo foi aplicada à séries temporais artificias, com características sazonais e caóticas, e séries reais relacionadas a dados fisiológicos, sendo estudado o parâmetro de medida de similaridade. Os resultados demostraram que não foi possível constatar diferença estatisticamente significativas entre as medidas estudas e aplicadas nos conjuntos de dados selecionados.

Introdução

O avanço tecnológico possibilitou o armazenamento de grandes quantidades de dados em diferentes formatos, sendo de interesse de diversas áreas a extração de conhecimento desses dados (HAN; KAMBER, 2006). Nesse sentido, a análise de informações que possuem uma ordem cronológica, principalmente sua representação como Séries Temporais (ST), têm despertado grande interesse para diversas aplicações como recuperação de conteúdo, classificação, descoberta de *motifs*, previsão, entre outras.

Neste trabalho é apresentada uma avaliação do algoritmo de Aprendizado de Máquina *k-Nearest Neighbor* (kNN) (HAN; KAMBER, 2006) para previsão de dados temporais, proposto por Ferrero (2009), denominado *k-Nearest Neighbor Time Series Prediction* (kNN-TSP).

Materiais e métodos

O algoritmo kNN-TSP é um método não linear para previsão de Séries Temporais proposto por Ferrero (2009) e que baseia-se na busca por subsequências passadas de morfologia similares para predizer valores futuros. Esse método é composto pelos seguintes passos:

- **Passo 1 – Dados de Entrada:** o algoritmo necessita como entrada de dados a ST que terá seus valores futuros estimados, o tamanho da janela deslizante (*w*), a função de previsão utilizada para o cálculo dos valores futuros e a quantidade de vizinhos mais próximos a serem considerados (*k*);
- **Passo 2 – Identificação de Subsequências Similares:** é utilizada uma medida de similaridade, para quantificar à similaridade, entre os *w* pontos finais da ST e todas as subsequências candidatas a vizinhos mais próximos;
- **Passo 3 – Seleção das *k* Subsequências:** com as subsequências quantificadas, estas são ordenadas de acordo com sua similaridade e são selecionadas as *k* subsequências mais similares;
- **Passo 4 – Cálculo da Função de Previsão:** com as *k* subsequências mais próximas obtidas da etapa anterior é realizado o cálculo do valor futuro, utilizando a função de previsão;
- **Etapa 5 – Dados Previstos:** os dados de saída são os valores futuros estimados da ST.

Para a implementação do algoritmo foi utilizado à linguagem de programação e o ambiente matemático e estatístico R versão 2.15 e ferramenta RStudio versão 0.95 para o desenvolvimento de códigos em R.

Na avaliação experimental foram selecionadas cinco séries temporais artificiais, pertencentes a duas famílias distintas: modelos sazonais e modelos caóticos. Estás séries são geradas a partir de modelos matemáticos descritos em (KULESH et al., 2008). Com o objetivo de avaliar o desempenho do algoritmo quando aplicado a dados reais foram selecionadas séries temporais disponíveis no repositório da *Physionet*[1].Portanto as ST reais selecionadas referem-se a dados de frequência cardíaca, de ruídos causados pela respiração humana e de concentração de oxigênio no sangue. Essas observações são pertencentes a um paciente com apneia do sono.

Após, foram definidos os parâmetros necessários para a execução do algoritmo, dentre os quais as medidas de similaridade Manhattan, Euclidiana e Métrica *L3*. Como função de previsão foi utilizada a média de valores relativos proposta por FERRERO (2009) e para o número de subsequências similares utilizou-se *k* variando de 1 à 5. Para as ST artificiais o parâmetro de tamanho

[1] http://physionet.org/physiobank/database/santa-fe/

das subsequências foi definido como o sugerido por Kulesh et al. (2008) e nas séries reais optou-se por utilizar 10% do tamanho das ST.
Na Tabela 1 são apresentadas as características e os parâmetros utilizados na avaliação.

Tabela 1 – Configuração do parâmetro *w* e número dos valores previstos

Séries Temporais Artificiais				
Identificador	**Série Temporal**	***m***	***w***	**Pontos Previstos**
(a.1)	Dependência sazonal	2200	100	220
(a.2)	Sazonalidade multiplicativa	590	15	88
(a.3)	Alta frequência	550	70	55
(b.1)	Lorenz	551	25	100
(b.2)	Mackey-Glass	551	7	100
Séries Temporais Reais				
(c.1)	Frequência cardíaca	900	90	90
(c.2)	Ruída causada pela respiração	900	90	90
(c.3)	Concentração de oxigênio no sangue	900	90	90

Os valores previstos foram avaliados por meio do *Mean Absolute Percentage Error* (MAPE). Os resultados foram analisados com o teste estatístico de Friedman para dados emparelhados e comparações múltiplas, com nível de significância de 5% ($p < 0,05$) e pós-teste de Dunn.

Resultados e Discussão

Nas Tabelas 2 e 3 são apresentado os valores de média e desvio padrão de MAPE para as séries temporais artificiais e reais, nessas tabelas é possível observar, que a distância de Manhattan obteve valores médios de erros menores que as demais medidas, com exceção da série temporal artificial (a.1) e série temporal real (c.2). No entanto, não houve diferença estatisticamente significativa entre nenhuma das comparações.

Tabela 2 – Valores de média e desvio padrão de MAPE para séries reais

Séries Temporais Reais				
Identificador		**Manhattan**	**Euclidiana**	**Métrica *L3***
(c.1)	Média (Desvio Padrão)	1,8900 (0,0928)	1,9437 (0,1233)	1,9873 (0,1500)
(c.2)	Média (Desvio Padrão)	99,2926 (13,2029)	90,9129 (7,6817)	109,6470 (11,4690)
(c.3)	Média (Desvio Padrão)	0,1569 (0,0231)	0,1480 (0,0244)	0,1477 (0,0193)

Tabela 3 – Valores de média e desvio padrão de MAPE para séries artificiais

Séries Temporais Artificiais				
Identificador		**Manhattan**	**Euclidiana**	**Métrica *L3***
(a.1)	Média (Desvio Padrão)	0,0022 (0,0014)	0,0022 (0,0014)	0,0022 (0,0014)
(a.2)	Média (Desvio Padrão)	16,1830 (4,8085)	16,1888 (8,8169)	16,3307 (5,0298)
(a.3)	Média (Desvio Padrão)	13,6437 (1,6770)	12,8901 (1,6855)	12,9581 (1,6823)
(b.1)	Média (Desvio Padrão)	42,6150 (15,7517)	46,0198 (13,5294)	48,8610 (11,2682)
(b.2)	Média (Desvio Padrão)	1,2966 (0,1268)	1,1900 (0,1421)	1,2362 (0,1547)

Conclusões

Neste trabalho foi realizada uma avaliação experimental do algoritmo kNN-TSP, considerando seus principais parâmetros quando aplicado tanto a dados artificiais quanto reais. Como resultado observou que não houve diferença estatisticamente significativa entre as diferentes variações de parâmetros. Trabalhos futuros incluem um estudo da influência do parâmetro de tamanho de subsequência e outras características do algoritmo.

Agradecimentos

Ao Programa Institucional de Bolsas de IC (PIBIC/Fundação Araucária/Unioeste) e ao LABI.

Referências

FERRERO, C. A. **Algoritmo kNN para previsão de dados temporais: funções de previsão e critérios de seleção de vizinhos próximos aplicados a variáveis ambientais em limnologia**. 2009. Dissertação (Mestrado) - Instituto de Ciências Matemáticas e Computação, Universidade de São Paulo, São Carlos, 2009.

HAN, J.; KAMBER, M. **Data Mining: Concepts and Techniques**. 2. ed. San Francisco: Morgan Kaufmann, 2006.

KULESH, M.; HOLSCHNEIDER, M.; KURENNAYA, K. Adaptive Metrics in the Nearest Neighbours Method. **Physica D: Nonlinear Phenomena**, v. 237, n. 3, p. 283–291, 2008.