http://dx.doi.org/10.5540/DINCON.2011.001.1.0087

DINCON 2011
10ª Conferência Brasileira de Dinâmica, Controle e Aplicações
28 de agosto a 1º de setembro de 2011

ESTUDO DA INFLUÊNCIA DE MEDIDAS DE SIMILARIDADE DA NORMA L_P NO ALGORITMO *kNN-TSP* PARA PREVISÃO DE DADOS TEMPORAIS

Jorge Aikes Junior[1], *Huei Diana Lee*[1,2], *Carlos Andrés Ferrero*[1], *Willian Zalewski*[1], *Feng Chung Wu*[1,2]

[1]Laboratório de Bioinformática – LABI – Universidade Estadual do Oeste do Paraná, Foz do Iguaçu, Brasil
jorge_a_junior@hotmail.com, {hueidianalee, anfer86, willzal, wufengchung}@gmail.com
[2]Programa de Pós-Graduação em Ciências da Cirurgia – Faculdade de Ciências Médicas – Universidade Estadual de Campinas, Campinas, Brasil

Resumo: Neste trabalho é realizado um estudo sobre a influência das medidas de similaridade da Norma L_p na previsão de dados em séries temporais. Utiliza-se como estratégias de previsão a busca de padrões similares do passado para a previsão de eventos futuros com uma adaptação do algoritmo *k-Nearest Neighbor*.

Palavras-chave: Análise de Séries Temporais, Aprendizado de máquina, Modelo não-paramétrico.

1. INTRODUÇÃO

O advento da tecnologia tem motivado a utilização de sistemas computacionais para a aquisição e o gerenciamento de dados em diversas áreas. Dependendo do objetivo, esses sistemas possibilitam o armazenamento de grande volume de dados em diferentes formatos. A realização de uma análise não-automática, com o intuito de extrair padrões relevantes, pode ser uma tarefa demorada e sujeita à subjetividade, tornando-se, em alguns casos, inviável devido à alta complexidade da relação entre os dados. Nesse sentido, entre outros esforços, métodos e ferramentas computacionais têm sido desenvolvidos para auxiliar na análise mais completa dessas informações[1]. A área de Mineração de Dados (MD) tem como principal objetivo desenvolver métodos e ferramentas para a extração de conhecimento em Bases de Dados (BD) [1]. No entanto, os métodos tradicionais de MD para a construção de modelos não levam em consideração a característica temporal na análise dos dados. Desse modo, pesquisas têm sido desenvolvidas propondo a adaptação desses métodos de análise, nos quais o tempo constitui um fator importante [2, 3].

Dentre as tarefas de interesse existentes, destaca-se a previsão de dados em Séries Temporais (ST), a qual tem como objetivo a estimativa de dados desconhecidos a partir de um conjunto de informações conhecidas. Essa tarefa é de interesse em múltiplas áreas do conhecimento, por exemplo: em economia [4]; em hidrologia [5]; e em medicina [6].

Uma das abordagens para a previsão de dados consiste na busca por padrões similares no passado para a previsão de eventos futuros. Nessa abordagem, um dos parâmetros que influencia na precisão consiste na medida de similaridade utilizada para identificar as sequências similares na série, as quais são utilizadas posteriormente para estimar valores futuros. Dentre as medidas amplamente divulgadas na literatura têm-se as baseadas na Norma L_p [2, 7, 8].

Neste trabalho é apresentado um estudo preliminar da influência de medidas de similaridade da Norma L_p na previsão de dados em séries temporais artificiais de modelos sazonais e de modelos caóticos, e em séries reais referentes ao fluxo de transporte rodoviário, considerando o algoritmo *k-Nearest Neighbor - Time Series Prediction* (*kNN-TSP*) [2] pertencente à abordagem anteriormente citada.

Este trabalho faz parte do projeto de Análise Inteligente de Dados em uma parceria entre o Laboratório de Bioinformática 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 do Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo (USP)/São Carlos e o Grupo Interdisciplinar em Mineração de Dados e Aplicações da Universidade Federal do ABC (UFABC).

O presente trabalho está organizado do seguinte modo: na Seção 2 são apresentados conceitos básicos sobre ST e previsão de valores, bem como o algoritmo *kNN-TSP*; na Seção 3 são descritos os dados e a configuração experimental utilizada com o algoritmo; na Seção 4 é apresentada a avaliação experimental e a discussão dos resultados; e, na Seção 5 são apresentados a conclusão e os trabalhos futuros.

[1]Neste trabalho os termos dado e informação são usados indistintamente.

2. SÉRIES TEMPORAIS E PREVISÃO DE VALORES

As ST podem ser entendidas como um conjunto de observações ordenadas no tempo. Assim, pode-se definir uma ST Z de tamanho m como uma série ordenada de observações, ou seja, $Z = (z_1, z_2, \ldots, z_m)$ em que $z_t \in \mathbb{R}$, sendo que z_t é uma observação no instante de tempo t [9].

A previsão em ST busca, por meio da exploração de dados conhecidos, projetar dados futuros. De modo geral, as abordagens de previsão podem ser classificadas em paramétricas e não-paramétricas. Os métodos paramétricos assumem que os dados respeitam alguma distribuição e que podem ser modelados a partir de um conjunto de parâmetros. Entre os principais modelos paramétricos podem ser citados: Auto-regressivos (AR), Médias Móveis (MA), Auto-regressivos de Médias Móveis (ARMA) e Auto-regressivos de Médias Móveis Integrados (ARIMA) [9]. Os métodos não-paramétricos não definem parâmetros de uma distribuição específica e têm a capacidade de se adaptar a diferentes comportamentos ao longo do tempo. Entre os métodos propostos encontram-se os baseados nos conceitos de redes neuronais artificiais e variações do algoritmo dos vizinhos mais próximos ou *k-Nearest Neighbor* (*kNN*) [2].

2.1. Algoritmo kNN-TSP

Em [2], é proposta uma adaptação do algoritmo *kNN* para a previsão de valores em ST denominado *k-Nearest Neighbor - Time Series Prediction* (*kNN-TSP*). Esse algoritmo consiste em encontrar as k sequências de tamanho w mais similares na porção conhecida da ST, e com base nas informações dessas sequências, e em uma função de previsão, realizar a estimativa do valor futuro.

A previsão com o algoritmo *kNN-TSP* depende de alguns parâmetros: **1 - Tamanho *w* da janela para extrair as sequências** — refere-se ao tamanho das sequências a serem consideradas para o cálculo do valor futuro na ST; **2 - Conjunto de exemplos de treinamento** — consiste no conjunto de sequências pertencentes à ST a serem consideradas para constituir o conjunto de treinamento; **3 - Medida de similaridade** — é utilizada para quantificar a similaridade entre os exemplos; **4 - Cardinalidade do conjunto de sequências similares** — refere-se à quantidade (k) de sequências mais próximas a serem consideradas para a previsão do valor futuro e **5 - Função de previsão** — é utilizada para determinar a maneira como serão considerados os valores das sequências mais próximas para estimar o valor futuro.

2.2. Medidas de Similaridade e a Norma L_p

A medida de similaridade define o critério para quantificar quão similares são duas sequências e decidir se pertencem ou não a um determinado padrão. Nesse sentido, diversas técnicas têm sido propostas na literatura e avaliadas em função de alguma tarefa de interesse: em [8] é proposta a combinação de medidas de similaridade para desenvolver uma alternativa à distância Euclidiana; em [3], medidas foram avaliadas para o reconhecimento de ST de mercado financeiro; no trabalho de [10], medidas são comparadas para detecção de anomalias em ST; e em [11] é proposta uma métrica adaptativa para melhorar a precisão de previsão de ST.

O algoritmo *kNN-TSP* foi desenvolvido utilizando a distância Euclidiana como medida para definir a similaridade entre as sequências [2]. Essa distância pertence a um conjunto de medidas, conhecidas como Norma L_p, as quais são amplamente divulgadas e utilizadas na literatura [2, 7, 8].

Para o cálculo da distância baseado nessa norma, cada sequência é considerada um ponto no espaço w-dimensional. Desse modo, a similaridade entre essas sequências é dada pela diferença entre esses pontos (Equação 1) [7]:

$$L_p(x,y) = \left(\sum_{i=1}^{w} |x_i - y_i|^p \right)^{\frac{1}{p}} \quad (1)$$

em que x e y são vetores w-dimensionais, os quais representam sequências contidas em uma ST; e p define a medida de distância a ser utilizada. Quando $p = 1$ a medida é chamada de distância Manhattan, ou *City Block* (L_1); quando $p = 2$ tem-se a distância Euclidiana (L_2) e com $p = 3$ a distância pode ser denominada Métrica L_3 (L_3). Deve ser observado que, à medida que o valor de p aumenta, a quantidade de operações a serem realizadas também aumenta, o que eleva o custo computacional.

3. CONFIGURAÇÃO EXPERIMENTAL

Para estudar a influência das medidas de similaridade na precisão de previsão do algoritmo *kNN-TSP*, foram utilizadas cinco ST artificiais [11], agrupadas em duas famílias, de acordo com as suas características: séries temporais de modelos sazonais (a); e séries temporais de modelos caóticos (b). Foram utilizadas também duas ST reais disponibilizadas pela edição do ano de 2010 da *Time Series Forecasting Grand Competition for Computational Intelligence (NN GC1)*[2]. As séries são apresentadas a seguir.

3.1. Modelos Sazonais (a)

Essas ST permitem avaliar algoritmos de previsão em comportamentos previsíveis. Em geral, apresentam tendência definida e mudança de amplitude de modo sazonal [11]. Foram utilizadas três séries: **ST de dependência sazonal (a.1)** — possui sazonalidade constante e tendência linear, **ST de sazonalidade multiplicativa (a.2)** — considera variação de tendência não-linear e sazonalidade multiplicativa — e **ST de alta frequência (a.3)** — possui dados que consideram sazonalidade multiplicativa e suave aumento de amplitude.

3.2. Modelos Caóticos (b)

Essas ST permitem avaliar algoritmos de previsão, frente a comportamentos pouco previsíveis e que apresentam ciclos não-repetitivos. Esse fato torna essas séries menos previsíveis do que as anteriores [11]. Foram utilizadas duas séries: **ST de Lorenz (b.1)** — construída por meio de um sistema de equações diferenciais — e **ST de Mackey-Glass (b.2)** — construída por meio de um sistema de equações originalmente criado para modelar a formação de linfócitos.

3.3. Séries temporais reais

As ST reais foram disponibilizadas pela *NN GC1*, que é uma competição de previsão de ST com o objetivo de avaliar

[2] http://www.neural-forecasting-competition.com/

a acurácia de métodos de Inteligência Computacional aplicados à previsão de ST. Estão disponíveis seis BD, cada uma contendo onze ST, relacionadas à tranporte, com diferentes intervalos de aquisição. Dentre as ST disponíveis foram escolhidas as séries (1.E-001) e (1.E-009) (Figura 1), por possuírem o menor e o maior número de observações, respectivamente, da BD "NNG-E", com frequência diária. Essas séries representam a quantidade de veículos dirigindo-se em apenas uma direção e a quantidade de veículos deslocados em ambas as direções em túneis na Suiça.

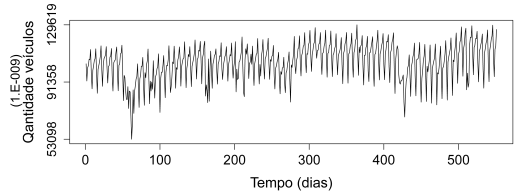

Figura 1 – ST (1.E-009) (*NN GC1*).

3.4. Definição dos Parâmetros do Algoritmo kNN-TSP

O valor do tamanho de janela (w) para as séries artificiais, foi selecionado baseado em [11] e para as séries reais foi definido com base na sazonalidade sugerida por [12]. Com relação à quantidade de valores previstos das séries reais, foi utilizada a mesma quantidade indicada pela organização da competição para essas bases. Na Tabela 1 são apresentados o tamanho da série (m); os valores de w e o número de valores a serem previstos em cada ST. Em relação ao parâmetro de medida de similaridade foram utilizadas as medidas da Norma L_p, com valores de $p = 1$, 2 e 3. Os valores de k foram avaliados de $k = 1$ a 5 e como função de previsão foi utilizada a Média de Valores Relativos [2, 5].

Tabela 1 – Descrição das ST e parâmetros.

Id	Série Temporal	*m*	*w*	Val. 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
(1.E-001)	Vans em uma direção	377	7	14
(1.E-009)	Veículos em ambas direções	747	7	14

Os valores previstos foram avaliados por meio do Erro Médio Absoluto (EMA) [13], entre cada valor previsto e o valor pertencente à série original. A análise das previsões foi realizada individualmente para cada ST. 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 [14].

4. RESULTADOS E DISCUSSÃO

Na Figura 2 são apresentados os valores de EMA para as séries (b.2) e (1.E-009). Os gráficos das demais séries não foram apresentados por questão de espaço, sendo escolhidos esses por representarem o comportamento geral das demais séries. Em cada gráfico, o eixo das abscissas indica o número de vizinhos próximos; o eixo das ordenadas o valor de EMA[3]; e os pontos, na forma de círculo vazio, triângulo e círculo preenchido, indicam os valores de EMA para as medidas de similaridade da Norma L_p, para $p = 1$, 2 e 3, respectivamente.

Na análise dos gráficos, as séries (a.2) e (b.1) apresentaram comportamento semelhante à série (b.2) (Figura 2(b.2)), onde, em geral, valores próximos a $k = 1$ apresentaram menor EMA e valores próximos de $k = 5$ mostram maior EMA. Além disso, observa-se leve tendência crescente do EMA à medida que aumenta o número de vizinhos próximos, para quaisquer das medidas utilizadas. Já para as séries (a.1) e (a.3) não foi possível observar essa característica. Nas séries (1.E-001) e (1.E-009) é possível observar que valores próximos de $k = 5$ apresentam um EMA menor que valores próximos de $k = 1$. Este comportamento apresenta-se contrário ao das ST artificiais, reafirmando que o valor de k indicado é dependente da característica de cada série. Observa-se também, para série (1.E-001), uma tendência de decrescimento do EMA à medida que aumenta o valor de k.

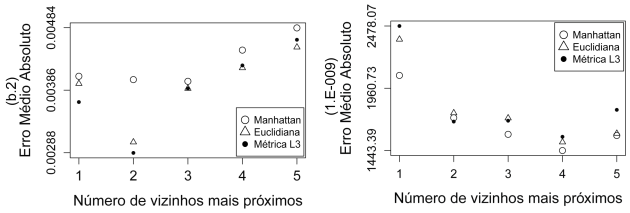

Figura 2 – EMA para as séries (b.2) e (1.E-009).

Na análise estatística dos resultados, para as séries (a.1), (a.2), (b.1), (b.2), (1.E-001) e (1.E-009) não foi constatada diferença estatisticamente significativa (**d.e.s**) entre os erros de previsão das medidas de similaridade utilizadas. Já a série (a.3), apresentou **d.e.s** entre as distâncias Manhattan e Métrica L_3 ($p < 0,01$), sendo que entre as demais distâncias não foi evidenciada **d.e.s**.

Foi também realizado um estudo comparativo (Tabela 2), para cada uma das ST e cada uma das três medidas de similaridade, com o intuito de analisar quão bem os valores previstos se relacionam com os valores observados, utilizando o Coeficiente de Determinação (R^2) [14].

Tabela 2 – Coeficientes de Determinação.

Id	Manhattan	Euclidiana	Métrica L_3
(a.1)	1,00000	1,00000	1,00000
(a.2)	0,91909	0,91908	0,91402
(a.3)	0,97353	0,96896	0,96931
(b.1)	0,99728	0,99713	0,99698
(b.2)	0,99895	0,99912	0,99912
(1.E-001)	0,68760	0,76955	0,74802
(1.E-009)	0,94656	0,94677	0,94534

Analisando-se a Tabela 2 percebe-se que, para todas as séries artificiais, exceto (b.2), a distância Manhattan apresentou os maiores valores de R^2. Especificamente para a série (a.1), a aplicação do Coeficiente de Determinação apresentou o valor de $R^2 = 1,00000$ para as três medidas avaliadas, o que

[3] O eixo das ordenadas dos gráficos encontra-se em escala diferente devido às ST apresentarem comportamentos variados e em intervalos de valores distintos.

mostra a alta relação entre os valores observados e os previstos, independentemente da medida utilizada. Para a série (a.2) observa-se uma tendência decrescente do R^2 à medida que aumenta o valor de p. Para as séries (a.3) e (b.2), os resultados de R^2 apresentaram-se ligeiramente diferentes entre as três medidas, porém, não exibem uma tendência de crescimento definida de acordo com a norma utilizada. Os valores de R^2 para a série (b.1) estão muito próximos, indicando menor influência da medida de similaridade nos resultados das previsões. Para as séries (1.E-001) e (1.E-009) a distância Euclidiana apresentou o melhor resultado de R^2, porém sem uma tendência clara de melhora relacionada ao valor de p.

Em uma análise geral, a distância Manhattan parece ser, neste trabalho, a medida de similaridade mais adequada, pois, é a medida que apresenta menor custo computacional e para as séries (a.1), (a.2), (b.1), (b.2), (1.E-001) e (1.E- 009) não foi possível constatar **d.e.s**. Já na série (a.3) não houve predominância de uma medida com menor valor de EMA. Na análise conjunta dos valores de R^2 de todas as ST em relação a cada uma das medidas de similaridade não foi possível constatar **d.e.s**. Este resultado corrobora com os resultados apresentados pela análise do EMA, indicando que, para a maioria dos casos, a distância Manhattan pode se tornar uma alternativa interessante, por apresentar resultados globais na previsão de valores de ST semelhantes às demais medidas avaliadas, porém com menor custo computacional.

5. CONCLUSÃO

Neste trabalho foi apresentado um estudo preliminar da influência de medidas de similaridade da Norma L_p para a previsão de valores em ST utilizando o algoritmo *kNN-TSP*.

De modo geral, a distância Manhattan apresentou-se, neste trabalho, como a medida de similaridade mais adequada, pois não foi constatada **d.e.s** entre as três medidas avaliadas, exceto para a série (a.3).

A partir da análise de R^2 foi possível verificar que, para as séries (a.1), (a.2) e (b.1), houve um decréscimo do valor do R^2 entre os valores observados e os previstos, à medida que aumenta o valor de p, referente à medida de similaridade, de um a três. Para séries com essas características, a utilização da distância Manhattan torna-se vantajosa devido ao alto valor de R^2 e menor custo computacional, se comparada às demais medidas avaliadas. As séries de (a.3), (b.2), (1.E-001) e (1.E-009) não apresentaram tendência de decrescimento definida relacionada ao valor de p. No entanto, devido aos valores de R^2 da distância Manhattan serem próximos aos das demais medidas, e ao menor custo computacional, essa medida pode ser considerada a mais apropriada para ser aplicada.

Trabalhos futuros incluem o estudo da influência de outros valores de p da Norma L_p e de outras medidas de similaridade na previsão de dados temporais com o algoritmo *kNN-TSP*; a combinação de medidas de similaridade para o cálculo de distâncias; o estudo da influência do número de vizinhos próximos na previsão de dados; e a utilização da abordagem desse trabalho em outras séries temporais reais.

AGRADECIMENTOS

À Fundação Parque Tecnológico Itaipu — FPTI-BR — pelo apoio por meio da linha de financiamento de bolsas.

REFERÊNCIAS

[1] J. Han and M. Kamber, *Data Mining:Concepts and Techniques*, 2nd ed. MKP, San Francisco, 2006.

[2] C. A. Ferrero, M. C. Monard, H. D. Lee, and F. C. Wu, "Proposta de uma Função de Previsão de Dados Temporais para o Algoritmo dos Vizinhos mais Próximos," in *Anais do XXXV CLEI*, pp. 1–10, Pelotas, 2009.

[3] Y. Ding, X. Yang, and J. Li, "A new representation and distance measure for financial time series," in *Proc. of the ICIFE*. IEEE, pp. 220–224, Chongqing, 2010.

[4] M. P. Hanias and P. G. Curtis, "Time Series Prediction of Dollar\Euro Exchange Rate Index," *Intl. Research Jour. of Finance and Economics*, No. 15, pp. 232–239, 2008.

[5] F. K. Odan, C. A. Ferrero., L. F. R. Reis, and M. C. Monard, "Análise Comparativa dos Modelos kNN-TSP e Série de Fourier para Previsão de Demanda Horária para Abastecimento de Água," in *Anais do XVIII ABRH*, pp. 1–30, Campo Grande, 2009.

[6] T. Verplancke, S. Van Looy, K. Steurbaut, D. Benoit, F. De Turck, G. De Moor, and J. Decruyenaere, "A novel time series analysis approach for prediction of dialysis in critically ill patients using echo-state networks." *BMC medical informatics and decision making*, Vol. 10, p. 4, 2010.

[7] C. C. Aggarwal, A. Hinneburg, and D. A. Keim, "On the surprising behavior of distance metrics in high dimensional spaces," in *Proc. of the 8th ICDT*. Springer-Verlag, pp. 420–434, London, 2001.

[8] F. Fabris, I. Drago, and F. Varejão, "A Multi-measure Nearest Neighbor Algorithm for Time Series Classification," in *Proc. of the 11th IBERAMIA*. Springer-Verlag, pp. 153–162, Lisboa, 2008.

[9] P. A. Morettin and C. M. C. Toloi., *Análise de Séries Temporais*, 2nd ed. Edgard Blücher, São Paulo, 2006.

[10] Q. Yan, S. Xia, and Y. Shi, "An anomaly detection approach based on symbolic similarity," in *Proc. of the CCDC*. IEEE, pp. 3003–3008, Xuzhou, 2010.

[11] M. Kulesh, M. Holschneider, and K. Kurennaya, "Adaptive metrics in the nearest neighbours method," *Physica D: Nonlinear Phenomena*, Vol. 237, No. 3, pp. 283–291, 2008.

[12] C. Lemke and B. Gabrys, "Meta-learning for time series forecasting in the NN GC1 competition," in *Proc. of the WCCI 2010*. IEEE, pp. 1–5, Barcelona, 2010.

[13] R. Hyndman and A. Koehler, "Another look at measures of forecast accuracy," *Intl. Jour. of Forecasting*, Vol. 22, No. 4, pp. 679–688, 2006.

[14] H. J. Motulsky, "GraphPad InStat 3.0 User's Guide," San Diego, 1995, http://www.graphpad.com.