1.03.02 – Ciência da Computação / Matemática da Computação

ANÁLISE DE SINAIS COM DISTÂNCIAS ESTOCÁSTICAS E DIFERENÇAS DE ENTROPIAS: FERRAMENTAS PARA ANÁLISE DE SÉRIES TEMPORAIS

Eduarda T. C. Chagas[1], Alejandro C. Frery[2]
1. Estudante de IC de Ciência da Computação, Ufal
2. Pesquisador do Laboratório de Computação Científica e Análise Numérica, Ufal

Resumo:

Este trabalho relata o processo de desenvolvimento de uma plataforma de análise dos descritores causais de uma série temporal oriundos da Teoria da Informação. A plataforma visa facilitar a análise dessas séries nos mais variados ramos da ciência, como por exemplo, a discriminação entre fenômenos estocásticos e caóticos [1], a identificação de padrões de comportamento em redes veiculares [2], a classificação e verificação de assinaturas *online* [3], na análise da robustez de redes [4], e a classificação de padrões de consumo de energia elétrica [5]. O sistema foi implementado na linguagem de programação *R* que além de fornecer ferramentas gráficas, também possui uma grande precisão numérica. Ambas as características de extrema importância ao longo deste trabalho. Após comentar brevemente a respeito de conceitos da Teoria da Informação necessários no processo de análise e modelagem de uma série temporal, expomos os resultados alcançados no decorrer do projeto e sugestões para futuros trabalhos.

Palavras-chave: Teoria da Informação, plataforma *R*, Estatística Computacional.

Apoio financeiro: CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico).

Introdução:

Séries temporais são conjuntos de dados obtidos a partir de um processo observacional ao longo de um determinado período de tempo, não necessariamente dividido em espaços iguais, sendo caracterizadas pela dependência serial existente entre as observações.

O estudo de séries temporais é tipicamente dividido em duas vertentes [6], a análise do domínio do tempo e do domínio da frequência, sendo utilizadas em ambas as abordagens os dados que resultam diretamente das observações coletadas, que por sua vez estão sujeitos a efeitos danosos de diversos tipos de contaminação. Uma solução alternativa, presente na literatura, para evitar os efeitos dessa contaminação, consiste no uso de métodos não paramétricos.

Há diversas ferramentas que auxiliam na análise clássica de séries temporais; para a plataforma *R*, existindo diversas bibliotecas para essa finalidade (ver https://cran.rproject.org/web/views/TimeSeries.html). Além destas opções, o usuário também pode contar com os softwares de visualização de séries temporais. No entanto, são limitadas as opções de bibliotecas e softwares que trabalham exclusivamente com técnicas não paramétricas.

O projeto aqui relatado tomou como ponto de partida a identificação das necessidades dos pesquisadores: uma ferramenta gráfica amigável e funcionalidades rápidas, eficientes e numericamente confiáveis. Outro requisito foi o da portabilidade para diversos sistemas operacionais e arquiteturas de hardware, e o uso de ferramentas FLOSS (*Free/Libre Open Source Software*).

Apresentamos, assim, o desenvolvimento de uma ferramenta portável, rápida e de boa qualidade numérica que possibilita análises interativas e exploratórias dos dados de uma série temporal através de técnicas provenientes da Teoria da Informação. Com ela, o usuário dispõe de um conjunto técnicas de análise presentes na literatura para processar e examinar seus dados de modo eficiente e com um mínimo período de aprendizado. A ferramenta é extensível.

Metodologia:

A primeira parte do projeto consistiu da apropriação do referencial teórico. Seja a série temporal $x = (x_1, x_2, \ldots, x_n)$. Ao invés de analisarmos os valores, transformaremos grupos de N valores (não necessariamente adjacentes) e padrões ordinais, e analisaremos a sua distribuição de frequência. Por exemplo, e sem perda de generalidade, com $N=3$ e para qualquer i viável, se $x_i < x_{i+1} < x_{i+2}$ assinaremos a esta tripla o padrão π_0 ; caso $x_i < x_{i+1} < x_{i+2}$ o padrão será π_1 e assim por diante. Com isso, há $N!$ possíveis padrões. Esta é conhecida como *simbolização de Bandt & Pompe* [7].

Forma-se, então, um histograma e, a partir dele, extraem-se quantificadores como, por exemplo, entropia,

distância estocástica a uma distribuição de equilíbrio, e complexidade estatística.

Seja, assim, $h=(h_1,...,h_{N!})$ o histograma de proporções dos $N!$ padrões observados a partir da série temporal x. Calculamos a entropia de Shannon

$$H(\boldsymbol{h}) = \sum_{i=1}^{N!} (-\log h_i) h_i, \qquad (1)$$

com a convenção - ∞ 0=0. A entropia de Shannon é o primeiro elemento a descrever a nossa série temporal. Ela mede a desordem do sistema que deu origem aos dados x.

Calculamos logo a distância de Jensen-Shannoon à distribuição uniforme $u=(1/N!,...,1/N!)$

$$D(\boldsymbol{h}, \boldsymbol{u}) = \sum_{i=1}^{N!} \left(h_i \log \frac{h_i}{u_i} + u_i \log \frac{u_i}{p_i} \right), \qquad (2)$$

em que $u_i = 1/N!$. Esta é uma medida de quão perto ou longe a dinâmica subjacente está de um processo sem informação nenhuma.

Finalmente, calculamos o terceiro descritor da nossa série temporal: a sua Complexidade Estatística:

$$C(\boldsymbol{h}, \boldsymbol{u}) = H(\boldsymbol{h}) D(\boldsymbol{h}, \boldsymbol{u}). \qquad (3)$$

Cada série temporal pode então ser descrita por um ponto (H(h), C(h, u)). O conjunto de todos os pares (H(h), C(h, u)) para qualquer série temporal descrita por padrões de comprimento N jaz em um subconjunto compacto R^2: o plano Entropia-Complexidade.

Embora aqui relatemos apenas o uso da entropia de Shannon e da distância de Jensen-Shannon, o sistema oferece outras entropias [8] e distâncias estocásticas [9]. Com essa contribuição do nosso sistema, as análises podem ser enriquecidas por outros descritores.

Durante o desenvolvimento deste trabalho foram estudadas diversas técnicas de análise de séries temporais, com foco nas ferramentas disponíveis na plataforma *R*. Após o período inicial de aprendizagem, seguido do levantamento dos requisitos do software, foi iniciada a implementação em *R*, usando o software livre de desenvolvimento integrado RStudio Desktop.

Resultados e Discussão:

Seguindo o modelo de engenharia de software em espiral, o sistema foi projetado e desenvolvido de forma modular, composto pelas seguintes unidades:

- Módulo de simbolização;
- Módulo de análise;
- Modulo de visualização e interação (em fase de desenvolvimento);

Esses módulos foram e estão sendo desenvolvidos seguindo um cronograma. Depois passaram pelas seguintes etapas:

- Integração dos módulos em um sistema;
- Teste e validação do sistema (em fase de desenvolvimento);
- Geração da interface gráfica (em fase de desenvolvimento).

Permite-se a leitura de dados em vários formatos (TXT, CSV ou XLSX), e o usuário a seguir poderá escolher:

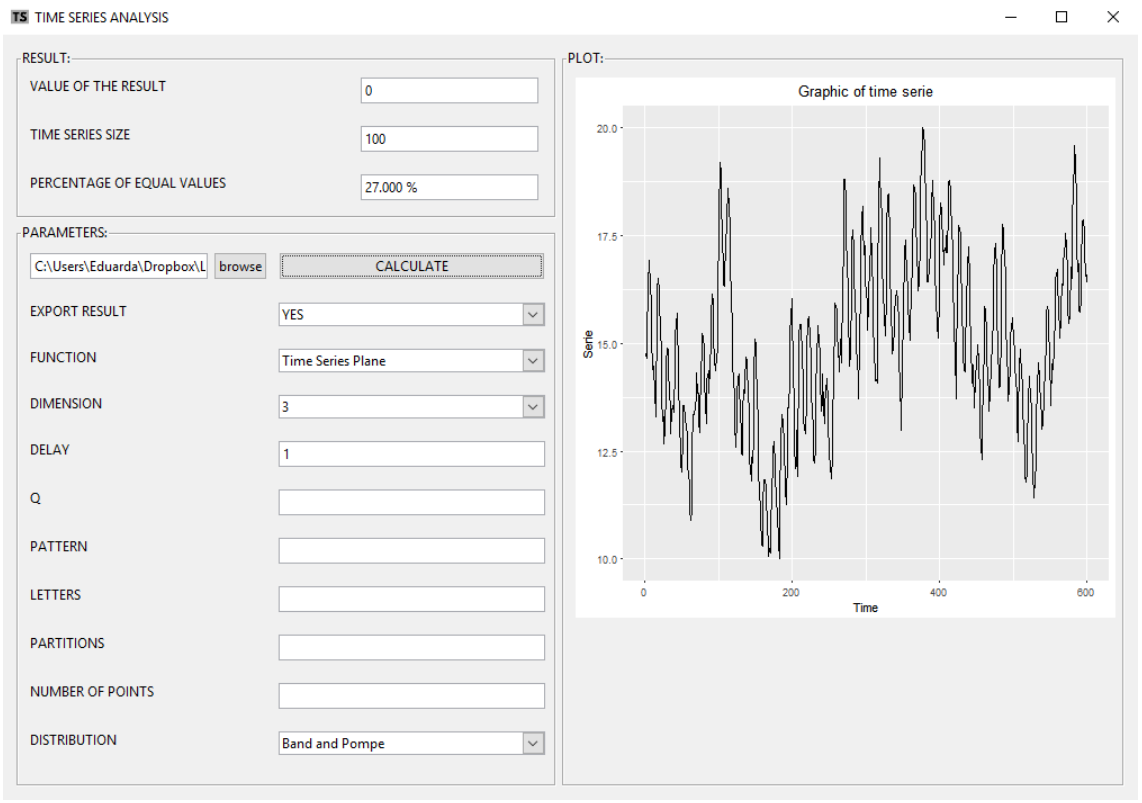
- Gerar o gráfico da série (ver Figura 1);
- Calcular seus diversos valores de Entropia;
- Calcular seus diversos valores de Distâncias estocásticas;
- Calcular complexidades estatísticas;
- Identificar padrões no gráfico da série temporal;
- Gerar planos de Entropias;

- Gerar planos de Distâncias estocásticas;
- Gerar o histograma de padrões;
- Identificar o ponto característico da série no plano Entropia-Complexidade.

Um elemento original do sistema é a vinculação entre o histograma de padrões e a série temporal. Escolhendo um ou mais elementos do histograma, os valores correspondentes na série temporal aparecem realçados. Esta funcionalidade permite a análise visual da distribuição temporal dos padrões, possibilitando futuramente a realização de outros testes.

O teste e a validação do sistema são tarefas contínuas, bem como o desenvolvimento de novas funcionalidades.

Figura 1. Imagem atual do software em processo de desenvolvimento.



Conclusões:

Através do desenvolvimento de tal plataforma por meio da linguagem *R*, fornecemos a base de geração de inúmeros outros modelos que tenham como objetivo a implementação de sistemas confiáveis que tornem mensuráveis as variadas propriedades presentes na teoria da informação, facilitando não apenas o estudo de séries temporais, como também todo o ramo atuante de análise de dados estatísticos.

Referências bibliográficas

[1] M. G. Ravetti, L. C. Carpi, B. A. Gonçalves, A. C. Frery, and O. A Rosso. **Distinguishing noise from chaos: objective versus subjective criteria using Horizontal Visibility Graph.** *PLOS ONE*, 9(9):1–15, 2014.

[2] A. L. L. Aquino, T. S. G. Cavalcante, E. S. Almeida, A. C. Frery, and O. A. Rosso. **Characterization of vehicle behavior with information theory.** *The European Physical Journal B: Condensed Matter and Complex Systems*, 88(10):257–269, Oct 2015.

[3] O. A. Rosso, R. Ospina, and A. C. Frery. **Classification and verification of handwritten signatures with time causal information theory quantifiers.** *PLOS ONE*, 11(12):e0166868, 2016.

[4] T. A. Schieber, L. Carpi, A. C. Frery, O. A. Rosso, P. M. Pardalos, and M. G. Ravetti. **Information theory perspective on network robustness.** *Physics Letters A*, 380:359–364, 2016.

[5] A. L. L. Aquino, H. S. Ramos, A. C. Frery, L. P. Viana, T. S. G. Cavalcante, and O. A. Rosso. **Characterization of electric load with information theory quantifiers**. *Physica A*, 465:277–284, 2017.

[6] P. J. Brockwell and R. A. Davis. ***Time Series: Theory and Methods*.** Springer-Verlag, Berlin, 2 edition, 1991.

[7] M. Salicrú, M. L. Mendéndez, and L. Pardo. **Asymptotic distribution of (h, ϕ)- entropy**. *Communications in Statistics – Theory Methods*, 22(7):2015–2031, 1993.

[8] L. Pardo. ***Statistical Inference Based on Divergence Measures***. Number 185 in Statistics, textbooks and monographs. Chapman & Hall/CRC, Boca Raton, 2006.

[9] C. Bandt and B. Pompe. **Permutation entropy: A natural complexity measure for time series**. *Physical Review Letters*, 88:174102–1–174102–4, Apr 2002.