Métricas para Séries Temporais como Núcleo de U-Estatísticas

Augusto Marcolin

Orientador: Marcio Valk

IM - UFRGS - Brasil

Resumo

O método de classificação e agrupamento de séries temporais baseado em U-estatísticas proposto por [1] tem como característica a dependência de uma medida de dissimilaridade entre séries temporais. Essas medidas são utilizadas como núcleo das U-estatísticas e suas características influenciam diretamente o comportamento da estatística de teste. Na literatura, existem uma grande variedade dessas medidas e o objetivo deste trabalho é realizar um estudo comparativo, através de simulações de monte carlo, para identificar qual medida é mais adequada para o método, considerando-se diferentes tipos de processos estacionários na configuração dos grupos.

Introdução

Atualmente existe uma demanda crescente pela utilização de métodos de classificação e agrupamento em séries temporais. Por esse motivo o assunto tem sido objeto de estudo em diversas áreas, tais como manutenção, medicina, biometria, química, astronomia,robótica, redes e indústria.

Nos métodos de agrupamento o objetivo é encontrar grupos de séries temporais que sejam similares dentro dos grupos e dissimilares entre os grupos. Essa tarefa não pode ser realizada com ferramentas convencionais, como a análise multivariada discreta, pois são ineficientes ou inapropriadas além de, geralmente, exigirem homoscedasticidade e normalidade dos dados para a inferência.

Método

A ideia do método é utilizar as medidas de dissimilaridade entre grupos e dentro dos grupos e mostrar que a estatística de teste composta pela diferença entre estas medidas é uma U-estatística e converge em distribuição para uma variável aleatória com distribuição normal.

Nesta seção apresentamos algumas das medidas de dissimilaridade mais comuns na literatura e que já estão implementadas no software R no pacote *"TSclust"*.

No domínio da frequência, a medida conhecida como *logaritmo do periodograma normalizado* (*DNLP*) é definida como a distância euclidiana entre os coeficientes dos periodogramas das séries x e y,

$$DLNP(x,y) = \frac{1}{T}\sum_{\ell=1}^{\lfloor \frac{T}{2} \rfloor} \left(I^*_x(\omega_\ell) - I^*_y(\omega_\ell)\right)^2,$$

Também no domínio da frequência, a medida de dissimilaridade entre duas séries temporais baseada na distância dos seus periodogramas integrados é definida por

$$INT.PER(x,y) = \int_{-\pi}^{\pi} |F_x(\lambda) - F_y(\lambda)| d\lambda$$

No domínio do tempo, uma das medidas é baseada na distância euclidiana ponderada entre os coeficientes de *autocorrelação*. O caso de ponderamento padrão será denotada por (*DAC*) e definida aqui por

$$DAC(x,y) = \sqrt{\sum_{h=1}^{L} \left(\widehat{\rho}_x(h) - \widehat{\rho}_y(h)\right)^2},$$

Igualmente relacionada a momentos amostrais, a medida de dissimilaridade baseada na correlação amostral (ou correlação de Pearson) é definida por

$$COR(x,y) = \sqrt{2(1-\rho)},$$

em que ρ denota a correlação de Pearson entre as séries x e y.

Uma medida adaptativa de dissimilaridade que cobre dissimilaridade no comportamento conjunto das séries e no comportamento dos coeficientes de correlação temporal é definida por

$$CORT(x,y) = \Phi[crt(x,y)]\delta(x,y),$$

Ainda no domínio do tempo, a medida que calcula a dissimilaridade baseada na distância euclidiana corrigida pela estimativa da complexidade da série é definida por

$$CID(x,y) = \delta(x,y) \times CF(x,y),$$

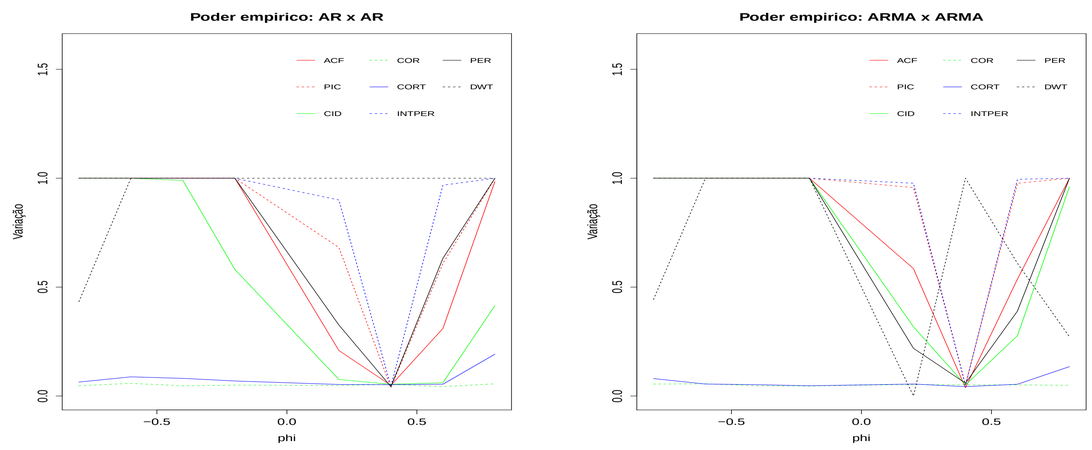
Outra maneira de medir dissimilaridade entre séries temporais é assumindo uma estrutura (modelo) para a série. Neste caso temos as chamadas *"model based distances"*. Assim, assumindo que a série pode ser representada através de um AR(∞), a medida baseada na distância euclidiana entre os coeficientes desta aproximação é chamada $AR.PIC(x,y)$. Ao substituir as séries temporais originais por seus coeficientes *"wavelets"* em uma escala apropriada e calcular a distância euclidiana entre esses coeficientes, temos a medida de dissimilaridade *DWT*.

Simulação

Realizamos um estudo de simulação para testar diferentes medidas de dissimilaridade para séries temporais, considerando primeiramente processos estacionários, em particular o processo autorregressivo (AR($\cdot$)). A primeira etapa consiste em gerar séries temporais artificiais a partir do processo AR(1). Para compor o primeiro grupo, foram geradas quatro séries a partir do processo AR(1), com coeficiente autorregressivo fixo $\phi_a = 0.4$, em que ϵ_t é aleatório com distribuição normal de média zero e variância um. Para compor o segundo grupo, foram geradas quatro séries a partir do mesmo processo X_t, mas com coeficiente autorregressivo tomando valores no conjunto $\phi_b \in \{-0.8, -0.6, -0.4, -0.2, 0.0, 0.2, 0.4, 0.6, 0.8\}$.

As séries do segundo grupo são geradas também a partir de um processo ARMA(1,1), com o mesmo coeficiente de média móvel $\theta = 0.5$, mas com coeficiente autorregressivo ϕ_b variando no conjunto $\{-0.8, -0.6, -0.4, -0.2, 0.2, 0.4, 0.6, 0.8\}$.

O tamanho de cada série considerada foi $T = 512$ e foram realizadas 1000 replicações para cada ϕ_b e as medidas de distâncias são as descritas anteriormente. Os resultados deste estudo podem ser observados na figura abaixo:



O método de classificação e agrupamento proposto por [1] depende diretamente da capacidade da medida de dissimilaridade diferenciar dois grupos distintos. Realizamos um estudo para testar a performance do método de classificação utilizando todas as medidas apresentadas anteriormente. O primeiro grupo de quatro séries é gerado a partir de um modelo AR(1) com coeficiente $\phi_a = 0.4$ e o segundo grupo com 4 séries é gerado a partir de um processo AR(1), mas com coeficiente autorregressivo $\phi_b \in \{-0.8, -0.6, -0.4, -0.2, 0.2, 0.4, 0.6, 0.8\}$. Uma série extra é gerada a partir do primeiro modelo e então classificada utilizando o método baseado em U-estatística com as métricas apresentadas anteriormente. Podemos observar na tabela abaixo o percentual de acerto para classificação de uma série temporal

ϕ_b	*ACF*	*PIC*	*CID*	*COR*	*CORT*	*INTPER*	*PER*	*DWT*
-0.8	100	100	100	52.1	63.2	100	100	100
-0.6	100	100	100	51.5	53.1	100	100	100
-0.4	100	100	100	49.2	50.3	99.9	99.9	100
-0.2	100	100	100	47.9	49.4	100	97.3	100
0.2	88.4	97.4	76	52.7	52	97.2	62	100
0.4	48.2	48.6	50	54.4	51.8	50.7	46.8	99.9
0.6	91.9	98	74.9	49.3	52.1	98.1	69.6	99.9
0.8	99.8	99.8	96.7	49.9	67.8	99.9	99.5	99.4

Bibliografia

- Valk, M. and A. Pinheiro (2012). Time-series clustering via quasi U-statistics. *Journal of Time Series Analysis*, vol.33(4), 608–619.
- Bagnall, A. and Janacek, G. (2005). Clustering Time Series with Clipped Data. *Machine Learning*. vol. 58, n. 2-3, pp. 151–178.
- Manso, P.M. (2013). *A package for stationary time series clustering*. Tese de Mestrado. Universidade da Coruña.

marciovalk@gmail.com