Universidade Federal do Rio Grande do Sul
Instituto de Matemática
Departamento de Estatística

Anais

IV SEMANÍSTICA

IV Semana Acadêmica do Departamento de Estatística

da UFRGS

http://www.ufrgs.br/semanistica

Porto Alegre - 20 a 22 de outubro de 2014

Medidas de dissimilaridade para o método de classificação de séries temporais baseado em U-estatísticas

Augusto Marcolin [1] [3]
Marcio Valk[2] [3]

Resumo:O método de classificação e agrupamento de séries temporais baseado em U-estatísticas 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.
Palavras-chave: *Séries temporais, Classificação, U-estatística.*

1 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. Na medicina, por exemplo, a série temporal pode ser de valores da pressão sanguínea de um paciente a cada hora, ou a taxa de batimentos cardíacos por minuto. Um dos objetivos desses métodos de classificação é reconhecer a qual grupo pré-determinado o sinal pertence. Na aplicação anterior esses grupos podem corresponder, por exemplo, ao estado de saúde de um paciente (pressão normal, alta ou baixa no sangue, ritmo cardíaco regular ou irregular).

O método de agrupamento proposto por [1] consiste na suposição de homogeneidade dos grupos, ou seja, sob H_0 o processo gerador das séries temporais é o mesmo para todos os grupos. A suposição essencial é que dentro de cada grupo temos homogeneidade. A ideia então é utilizar as medidas de dissimilaridade [1] 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

[1]UFRGS - Universidade Federal do Rio Grande do Sul. Email: *augustomarcolin@gmail.com*
[2]UFRGS - Universidade Federal do Rio Grande do Sul. Email: *marciovalk@gmail.com*
[3]

[1]O termo dissimilaridade é usado pois essas medidas não têm propriedade de uma distância

com distribuição normal. Essa convergência ocorre de duas formas: aumentando-se o tamanho das séries temporais e/ou aumentando-se o número de séries temporais (para mais detalhes ver [1]).

Na seção 2 apresentamos algumas das medidas de dissimilaridade mais comuns em séries temporais. Na seção 3 realizamos um estudo de simulação com o objetivo de verificar qual das medidas apresentadas na seção 2 têm melhor desempenho para uma determinada classe de processos estacionários, quando utilizadas como núcleo da estatística de teste no método proposto por [1].

2 Medidas de dissimilaridade entre séries temporais

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,

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

em que $I_x(\cdot)$ é a função periodograma da série x_t, $I^*{}_x(\omega) = \log\left[I_x(\omega)/\gamma_x(0)\right]$ é o logaritmo do periodograma normalizado e $\gamma_x(\cdot)$ é a função de autocovariância de x_t. Também no domínio da frequência, a medida de dissimilaridade entre duas séries temporais baseada na distância dos seus periodogramas integrados é definida por

$$INT.PER(x,y) = \int_{-\pi}^{\pi} |F_x(\lambda) - F_y(\lambda)| d\lambda \tag{2}$$

em que $F_x(\lambda_j) = C_x^{-1}\sum_{i=1}^{j} I_x(\lambda_i)$ e $F_y(\lambda_j) = C_y^{-1}\sum_{i=1}^{j} I_y(\lambda_i)$, com $C_x = \sum_{i=1} I_x(\lambda_i)$ e $C_y = \sum_{i=1} I_y(\lambda_i)$. Neste trabalho usamos a versão normalizada em que $C_x = C_y = 1$.

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

$$DAC(x,y) = \sqrt{\sum_{h=1}^{L}\left(\widehat{\rho}_x(h) - \widehat{\rho}_y(h)\right)^2}, \tag{3}$$

em que $\widehat{\rho}_x(h) = \widehat{\gamma}_x(h)/\widehat{\gamma}_x(0)$, é a função de autocorrelação de x_t e L o número de autocorrelações, que deve ser determinado de alguma forma. Neste trabalho usamos $L = 50$. Igualmente relacionada a momentos amostrais, a medida de dissimilaridade baseada na correlação amostral (ou correlação de Pearson) é definida por

$$COR(x,y) = \sqrt{2(1-\rho)}, \tag{4}$$

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

$$CORT(x,y) = \Phi[crt(x,y)]\delta(x,y), \tag{5}$$

em que $crt(x,y) = \sum_t (x_{t+1} - x_t)(y_{t+1} - y_t)(\sum_t (x_{t+1} - x_t)^2 \sum_t (y_{t+1} - y_t)^2)^{-\frac{1}{2}}$ é o coeficiente de correlação temporal de ordem um e mede a proximidade entre o comportamento dinâmico de x e y. A função $\Phi[u] = 2/(1+e^{ku})$, com $k \geq 0$ é chamada de *"adaptative tuning function"* e $\delta(x,y)$ é a distância euclidiana entre x e 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

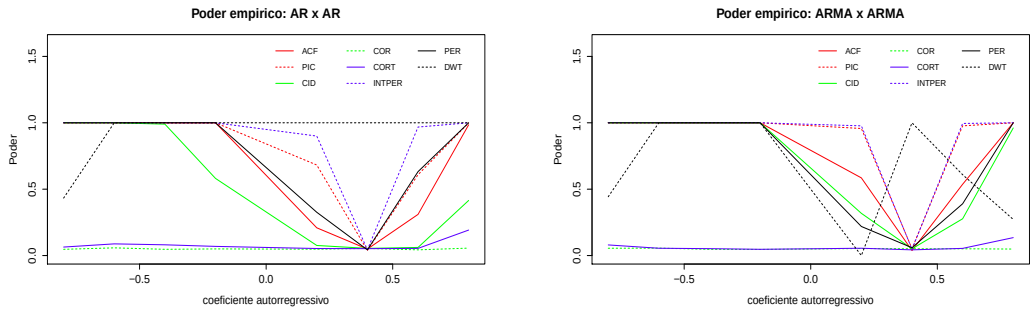
$$CID(x,y) = \delta(x,y) \times CF(x,y), \tag{6}$$

em que $CF(x,y)$ é o fator de correção de complexidade dado por $\max(CE(x), CE(y))/min(CE(x), CE(y))$, sendo $CE(x)$ a estimativa da complexidade de x definida por $CE(x) = \sqrt{\sum_{t=1}^{T}(x_{t+1} - x_t)^2}$. 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.

3 Estudo de 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 $\phi_b \in \{-0.8, -0.6, -0.4, -0.2, 0.0, 0.2, 0.4, 0.6, 0.8\}$. Com essa configuração, quando $\phi_a = \phi_b = 0.4$ tem-se a situação de homogeneidade entre os grupos (H_0), e, ao nível de 5% de significância, espera-se que o teste rejeite H_0 aproximadamente 5% das vezes. Isso dá uma estimativa do tamanho do teste. Para as demais combinações de ϕ_a e ϕ_b, espera-se que o teste rejeite a hipótese de homogeneidade dos grupos (H_0) na maioria das vezes. A proporção de vezes em que rejeita-se H_0 é uma estimativa do poder do teste. O tamanho de cada série considerada foi $T = 512$ e foram realizadas 1000 replicações para cada ϕ_b. As medidas de distâncias consideradas no estudo foram ACF, PIC, CID, COR, $CORT$, $INTPER$, PER e DWT, as quais são descritas na seção anterior.

Os resultados deste estudo podem ser observados na figura 1a. A medida com melhor desempenho foi $INTPER$, seguida de PIC, PER, ACF e CID, nesta ordem. Para as medidas COR e $CORT$, o teste não apresentou aumento no poder, mesmo em situações em que os modelos são muito diferentes, ou seja, em situações em que a diferença entre ϕ_a e ϕ_b é grande. No caso de DWT, aparentemente sua utilização não é adequada para este método. Um estudo semelhante foi realizado considerando-se agora quatro séries no primeiro grupo geradas a partir de um processo ARMA(1,1), com coeficientes autorregressivo e de média móvel fixos, a saber, $\phi_a = 0.4$ e $\theta = 0.5$. 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 $\{-0.8, -0.6, -0.4, -0.2, 0.2, 0.4, 0.6, 0.8\}$. Novamente, quando $\phi_a = \phi_b = 0.4$ tem-se a situação de homogeneidade entre os grupos (H_0), e, ao nível de 5% de significância, espera-se que o teste rejeite H_0 aproximadamente 5% das vezes. Os resultados exibidos na figura 1b mostram que neste caso as medidas $INTPER$ e PIC apresentam um desempenho equivalente e que a ACF é melhor que PER e CID. As medidas COR e $CORT$ continuam apresentando um fraco desempenho e DTW apresenta os mesmos problemas do caso anterior.



(a) AR x AR

(b) ARMA x ARMA

Figura 1: Estimativa do poder do teste para homogeneidade de dois grupos de séries temporais. No painel (a), o primeiro grupo contendo 4 séries é gerado a partir de um processo AR(1) com $\phi_a = 0.4$. As séries do segundo grupo são geradas também a partir de um processo AR(1), mas com coeficiente variando. O tamanho de cada série é $T = 512$. No painel (b), o primeiro grupo contendo 4 séries é gerado a partir de um processo ARMA(1,1) com coeficientes autorregressivo e de média móvel fixos em $\phi_a = 0.4$ e $\theta = 0.5$. As séries do segundo grupo são geradas também a partir de um processo ARMA(1,1), com $\theta = 0.5$ mas com coeficiente autorregressivo variando. Em ambos os casos, o tamanho de cada série é $T = 512$ e as métricas testadas foram ACF, PIC, CID, COR, $CORT$, $INTPER$, PER e DWT.

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 na seção 2. 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 na seção 2. Podemos observar na tabela 1 que exibe o percentual de acerto para classificação de uma série

Tabela 1: Percentual de acerto para classificação de uma série temporal. 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 na seção 2.

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

temporal, que as medidas que apresentaram melhor desempenho relativamente ao poder do teste, é que possuem uma capacidade maior de classificar corretamente, que é o caso da $INTPER$. O resultado da DWT não está correto, pois para $\phi_b = 0.4$, o percentual de acerto deveria ser aproximadamente 50%.

4 Conclusão

Neste trabalho, estudamos o comportamento de algumas medidas de dissimilaridades entre séries temporais quando utilizadas como núcleo da estatística de teste no método de classificação baseado em U-estatísticas. Podemos observar que, no caso em que as séries temporais advindas de processos estacionários, as medidas $INTPER$ e PIC apresentam melhor desempenho, relativamente ao poder do teste, nas diferentes configurações testadas. Isso reflete diretamente na capacidade do método classificar uma série temporal em seu respectivo grupo. Os próximos esforços são direcionados para busca de propriedades assintóticas da U-estatística de teste, quando o núcleo for uma dessas medidas que apresentaram melhor desempenho.

Referências

[1] Valk, M. and A. Pinheiro (2012). Time-series clustering via quasi U-statistics. *Journal of Time Series Analysis*, vol.33(4), 608–619.

[2] Bagnall, A. and Janacek, G. (2005). Clustering Time Series with Clipped Data. *Machine Learning*. vol. 58, n. 2-3, pp. 151–178.

[3] Manso, P.M. (2013). *A package for stationary time series clustering*. Tese de Mestrado. Universidade da Coruña.