Métricas para Séries Temporais como Núcleo de U-Estatísticas

Augusto Marcolin

Orientador: Marcio Valk

IM - UFRGS - Brasil

Resumo

A utilização de séries temporais está presente em diversos campos de estudo, nos quais observações ordenadas no tempo desempenham um importante papel na obtenção de resultados, tais como economia, biologia, etc. O método de classificação e agrupamento de séries temporais via U-estatísticas tem como objetivo comparar grupos de séries temporais, supondo que dentro de cada grupo todas as séries temporais têm o mesmo processo gerador, e concluir se o processo estocástico gerador de ambos os grupos é o mesmo. Muitas técnicas são apresentadas na literatura atual e todas dependem de medidas de distâncias entre séries temporais. Como existem na literatura enumeras medidas de distâncias para variáveis aleatórias em geral, pretendemos com esta pesquisa implementar essas distâncias para séries temporais e testar para diferentes classes de processos, a fim de determinar qual métrica se aplica melhor a cada tipo de série.

Introdução

Existem alguns trabalhos no sentido de identificar qual a melhor medida de distância entre duas séries temporais a ser usada em métodos de agrupamentos. Podemos citar os trabalhos de [8] no qual o autor apresenta algumas medidas de similaridades adaptadas para séries temporais e implementa-as no *software* R, e o trabalho de [7] em que o autor afirma que existem três tipos de objetivos quando o assunto é agrupamento de séries temporais. São eles:

- *Similaridade no tempo*. Por exemplo, se o objetivo é agrupar séries de preços de para descobrir mudança conjunta de preços, uma abordagem óbvia para medir a similaridade deste tipo de agrupamento é a utilização de uma métrica de distância baseado correlação ou a distância Euclidiana em dados normalizados.
- *Similaridade na forma*. Similaridades como tendências e mudanças estruturais que poderiam ser detectadas usando métricas baseadas em *dynamic time warping*
- *Similaridade em mudanças*. Por exemplo, um analista de ações poderia estar interessado em agrupar as ações que tendem a ter um aumento no preço da ação com uma queda no dia seguinte. A abordagem comum para este tipo de objetivo é assumir uma forma de modelo subjacente, tal como um modelo de oculto Markov.

No entanto, estes estudos utilizam basicamente métodos de agrupamento cujo resultado é empírico, ou seja, não são estatisticamente testáveis. Assim, se um destes métodos encontrar dois grupos distintos, não necessariamente significa que eles sejam gerados por dois processos distintos.

Nosso objetivo é utilizar a técnica proposta por [6] para agrupamento de séries temporais, o qual é capaz de testar se dois grupos têm o mesmo processo gerador, que também depende de medidas de distâncias entre as séries e realizar um estudo de simulação para verificar qual destas medidas é melhor para uma determinada classe de processos.

Método

Existem vários métodos para agrupar séries temporais e esses métodos consistem em encontrar grupos em que as séries sejam homogêneas dentro do grupo e heterogêneas entre os grupos, ou seja, a distância entre as series do grupo é a menor possível enquanto que a distância entre séries de grupos diferentes são maiores. Assim, é necessário medir distância entre séries temporais. Essa distância pode ser medida de diferentes formas:

- No domínio do tempo.
- No domínio da frequência.
- Medidas em que é necessário o ajuste de um modelo às séries.

Foram escolhidas oito medidas de distâncias para este trabalho, sendo duas delas no domínio da frequência(Periodograma Padronizado, Periodograma) e o restante no domínio do tempo:

. **Periodograma Padronizado**

$$d_{LNP}(X,Y) = \frac{1}{T}\sum_{\ell=1}^{\lfloor \frac{T}{2} \rfloor} \left(\bar{I}_X(\omega_\ell) - \bar{I}_Y(\omega_\ell)\right)^2,$$

em que $I_X(w_j) = \frac{1}{n}\left|\sum_{t=1}^{T} X_t e^{-itw_j}\right|^2$ é o periodograma de X na frequência de fourier w_j e $\bar{I}_X(\omega) = \log\left[\frac{I_X(\omega)}{\gamma_X(0)}\right]$ é o logaritmo do periodograma normalizado (LNP). Por esta razão, denotaremos esta medida por LNP.

. **Periodograma** - É a mesma relação do periodograma padronizado, porém é utilizado $I_X(w_j)$ ao invés de $\bar{I}_X(\omega)$

. **Autocorrelação**

$$d_{AC}(X,Y) = \frac{1}{T}\sum_{h=1}^{T}\left(\widehat{\rho}_x(h) - \widehat{\rho}_y(h)\right)^2,$$

em que $\widehat{\rho}_x(h) = \widehat{\gamma}_x(h)/\widehat{\gamma}_x(0)$, sendo

$$\widehat{\gamma}_x(h) = \frac{1}{T}\sum_{t=1}^{T-h}(X_t - \overline{X})(X_{t+h} - \overline{X}), \quad 0 \le h \le T-1$$

e L o número de autocorrelações, que deve ser determinado de alguma forma.

. **Esperança**

$$D_E(X,Y) = E\left(\frac{|X-Y|}{1+|X-Y|}\right).$$

. **Correlação**

$$D_C(Y,X) = \sqrt{2(1-\rho(Y,X)},$$

. **Euclidiana**

$$D_E(X,Y) = \sqrt{\sum_{t=1}^{T}(X_t - Y_t)^2}$$

. **Kolmogorov**
A métrica de Kolmogorov consiste em medir a distância entre as funções de distribuição acumulada. Seja F a função de distribuição de $\{X_t\}$ e G a função de distribuição de $\{Y_t\}$. Então

$$D_K(X,Y) = sup_x\left|F(x) - G(x)\right|.$$

Para a implementação, consideramos sua versão amostral

$$d_K(X,Y) = sup_x\left|\frac{1}{T}\left(\hat{F}(x) - \hat{G}(x)\right)\right|,$$

em que

$$\hat{F}(x) = \sum_{t=1}^{T} I(X_t), \text{ e } \hat{G}(x) = \sum_{t=1}^{T} I(Y_t),$$

com

$$I(X_t) = \begin{cases} 1, \text{se } X_t \in [a,x] \\ 0, \text{se } X_t \notin [a,x] \end{cases}$$

em que a é o menor de todos os X_t e Y_t.

. **Minkowski**

$$d_{MIK}(X,Y) = \left(\sum_{t=1}^{T}(X_t - Y_t)^q\right)^{\frac{1}{q}}$$

Todas estas medidas foram programadas utilizando o software R x64 2.13.2 . Após feitas as funções para o cálculo das distâncias, foi utilizado o método para comparar as séries temporais, que consiste em gerar, por exemplo, um grupo de séries a partir de um processo X e outro grupo a partir do processo Y, com $X \neq Y$ ou $X = Y$. Depois de geradas as séries, calculou-se as distâncias e então testou-se as hipóteses sob $H_0(X = Y)$,deve-se obter p-valor maior que 0.05 e sob $H_1(X \neq Y)$,deve-se obter p-valor menor que 0.05, também foi testada a normalidade na distribuição das distância, utilizando o teste de Kolmogorov-Smirnov.

Simulação

A primeira etapa foi gerar as séries temporais, o processo utilizado foi o AR(1). Geramos 4 séries do processo

$$X_t = \phi_a X_{t-1} + \epsilon_t,$$

e 4 séries do processo

$$Y_t = \phi_b Y_{t-1} + \epsilon_t,$$

com $-1 < \phi_a, \phi_b < 1$, $\phi_a \neq \phi_b$ e ϵ aleatório com distribuição normal com média zero e variância 1. As séries foram geradas para correlação 0, 0.3, 0.5, 0.8 e 0.99. Observamos até que ponto o teste rejeita que as séries provêm de um mesmo processo gerador. Queremos observar se há uma métrica mais adequada para este tipo de série temporal.
A primeira simulação foi realizada fixando $\phi_a = 0,2$ e $\phi_b = 0,8$, Obtemos os resultados que são apresentados na seguinte tabela:

AR(1) x AR(1), $\phi_1 = 0,2$ e $\phi_2 = 0,8$									
LNP					Period				
n	100	300	500	1000	n	100	300	500	1000
Correlação					Correlação				
0	0,01124	0,00428	0,00296	0,00464	0	0,38352	0,24440	0,25364	0,21856
0,3	0,00400	0,00240	0,00208	0,00240	0,3	0,14500	0,16408	0,12900	0,18360
0,5	0,02108	0,00256	0,00280	0,00188	0,5	0,13140	0,13484	0,12272	0,10256
0,8	0,00148	0,00188	0,00136	0,00160	0,8	0,01916	0,02412	0,01776	0,02092
0,99	0,00096	0,00084	0,00060	0,00120	0,99	0,00168	0,00144	0,00184	0,00196
Autocorrelação					Correlação				
n	100	300	500	1000	n	100	300	500	1000
Correlação					Correlação				
0	0,00472	0,00384	0,00312	0,00380	0	0,80520	0,82652	0,83288	0,82036
0,3	0,00348	0,00536	0,00240	0,00268	0,3	0,43468	0,44036	0,51120	0,46356
0,5	0,00096	0,00456	0,00200	0,00168	0,5	0,20276	0,26048	0,23524	0,23840
0,8	0,00120	0,00180	0,00124	0,00168	0,8	0,01736	0,01252	0,00768	0,01052
0,99	0,00160	0,00096	0,00148	0,00064	0,99	0,00172	0,00128	0,00176	0,00160
Esperança					Euclidiana				
n	100	300	500	1000	n	100	300	500	1000
Correlação					Correlação				
0	0,57688	0,54952	0,50408	0,51040	0	0,64632	0,69544	0,67692	0,65000
0,3	0,30816	0,27552	0,36320	0,33688	0,3	0,24648	0,19904	0,30128	0,25884
0,5	0,24352	0,23336	0,22392	0,20256	0,5	0,12664	0,09444	0,11676	0,10180
0,8	0,04116	0,02236	0,02156	0,02680	0,8	0,00752	0,00572	0,00460	0,00504
0,99	0,00200	0,00176	0,00208	0,00172	0,99	0,00176	0,00192	0,00168	0,00184
Kolmogorov					Minkowski				
n	100	300	500	1000	n	100	300	500	1000
Correlação					Correlação				
0	0,00896	0,00312	0,00140	0,00188	0	0,64408	0,73712	0,66720	0,66048
0,3	0,01660	0,00228	0,00168	0,00200	0,3	0,29444	0,22144	0,25736	0,23844
0,5	0,04280	0,00208	0,00204	0,00164	0,5	0,13496	0,07400	0,11744	0,10620
0,8	0,00224	0,00212	0,00264	0,00184	0,8	0,00480	0,00744	0,00592	0,00496
0,99	0,00164	0,00156	0,00176	0,00200	0,99	0,00152	0,00160	0,00200	0,00148

Pela tabela acima podemos notar que as distâncias que melhor se comportam são as da Autocorrelação, LNP e Kolmogorov, pois seus p.valores são mais baixos, sendo assim melhores em identificar diferenças entre grupos. Porém posteriormente foi testado a normalidade da estatística de teste Bn, que é baseada nas medidas de distâncias dentro dos grupos e entre os grupos, mais informações em [6]. Para a distância de Kolmogorov não foi aceita normalidade na estatística de teste, após notou-se que ela não é linearmente independente.

Bibliografia

- Caiado J., N. Crato and D. Peña (2006). A periodogram-based metric for time series classification. *Computational Statistics and Data Analysis*, vol. 50(10), 2668–2684.
- Galeano P. and Peña, D. (2001). Multivariate analysis in vector time seires. Resenhas, 383–404.
- Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *Ann. Math. Statistics*, vol. 19, 293–325.
- Pinheiro A., P.K. Sen and H.P. Pinheiro (2009). Decomposability of high-dimensional diversity measures: Quasi *U*-statistics, martingales and nonstandard asymptotics. *Journal of Multivariate Analysis*, vol. 100(8), 1645-1656.
- Pinheiro A., P.K. Sen and H.P. Pinheiro (2010). A class of asymptotically normal degenerate quasi *U*-Statistics. *Annals of the Institute of Statistical Mathematics*. DOI $10.1007/s10463 - 010 - 0271 - z$.
- 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