UMA APLICAÇÃO DE SISTEMAS DE CLASSIFICAÇÃO KNN PARA A IDENTIFICAÇÃO AUTOMÁTICA DE MODELOS ARMA SAZONAIS E NÃO SAZONAIS

Luiza Maria Oliveira da Silva
(Faculdades Ibmec - RJ)
luiza.maria@ibmecrj.br

Reinaldo Castro Souza
(PUC-Rio Pontifícia Universidade Católica do Rio de Janeiro - Departamento de Engenharia Elétrica)
reinaldo@ele.puc-rio.br

Maria Augusta Machado
(IBMECRJ)
mmachado@ibmecrj.br

Flávio Souza
(UERJ-Univers. do Estado do Rio de Janeiro)
flavioj@ism.com.br

Resumo

Este trabalho apresenta uma metodologia usando o classificador KNN para a identificação automática de estruturas Box & Jenkins sazonais e não-sazonais.

Palavras-chaves: Identificação de Estruturas Box & Jenkins, KNN, Classificação de dados.

Abstract

This paper presents a methodology using KNN classifier to identify automatically Box & Jenkins seasonals and not seasonals structures.

Key words : Box and Jenkins model identification, KNN, Classification.

1 - INTRODUÇÃO

Quando se quer realizar uma previsão de séries temporais, cogita-se sempre no uso da metodologia Box& Jenkins porque os resultados obtidos por este método, via de regra, mostram-se competitivos, em termos de desempenho, com os encontrados por quaisquer outros métodos.
Para identificar a estrutura Box & Jenkins mais adequada para a série em questão vários sistemas têm sido desenvolvidos.
O presente trabalho tem como objetivo apresentar uma metodologia que utiliza uma técnica de classificação de dados, conhecida como “K vizinhos mais próximos” (KNN – K Nearest Neighbors) para a identificação automática de Estruturas Box & Jenkins sazonais e não sazonais.

2. METODOLOGIA BOX & JENKINS

A análise de séries temporais, segundo Box & Jenkins (1976), tem como objetivo principal a realização de previsão. Essa metodologia permite que valores futuros de uma série sejam previstos tomando por base apenas seus valores presentes e passados. Isso é feito através da correlação temporal existente entre os valores existentes.

A realização do processo temporal pelo método de Box & Jenkins é representada por um conjunto de processos estocásticos denominados modelos ARIMA (*autoregressive integrated moving average*) onde em cada instante de tempo t, existe um conjunto de valores que a série pode assumir, aos quais estão associadas possibilidades de ocorrência.
Para cada instante de tempo t, é possível que exista uma função de densidade de probabilidade logo, cada variável aleatória $Z_t,\ t = t_1, t_2, \ldots$ pode ter média e variância específicas.
O trabalho consiste em descobrir qual é o processo que gera a série em estudo, isto é, qual o modelo que repesenta melhor a série.
Os modelos ARIMA simples de ordem (p, d, q) são definidos como

$$\phi(B)\ \nabla^d\ Z_t = \theta(B)\, a_t$$

e resultam da combinação de três componentes também denominados "filtros" que são:

- O componente auto-regressivo de ordem p – AR(p) (*autoregressive*);
- O filtro de integração de ordem d – I (*integrated*);
- O componente de médias móveis de ordem q – MA(q) (*moving average*).

Os modelos ARIMA sazonais de ordem $(P, D, Q)_s$ são definidos como

$$\Phi(B^s)\ \nabla_s^D\ \ Z_t = \Theta(B^s)\ \ a_t$$

e resultam da combinação de três componentes também denominados "filtros" que são:

- O componente auto-regressivo de ordem P – AR(P) (*autoregressive*);
- O filtro de integração de ordem D – I (*integrated*);
- O componente de médias móveis de ordem Q – MA(Q) (*moving average*).

Os modelos ARIMA multiplicativos de ordem (p, d, q) x $(P, D, Q)_s$ são aplicados a séries que apresentam correlação serial 'dentro' e 'entre' períodos sazonais e são definidos como

$$\Phi(B^s)\ \phi(B)\ \nabla_s^D\ \nabla^d\ Z_t = \Theta(B^s)\ \theta(B)\ a_t$$

A figura abaixo é uma representação em diagramas do modelo ARIMA (p, d, q) x $(P, D, Q)_s$ que ilustra a seqüência de filtragens aplicadas ao ruído branco (a_t).

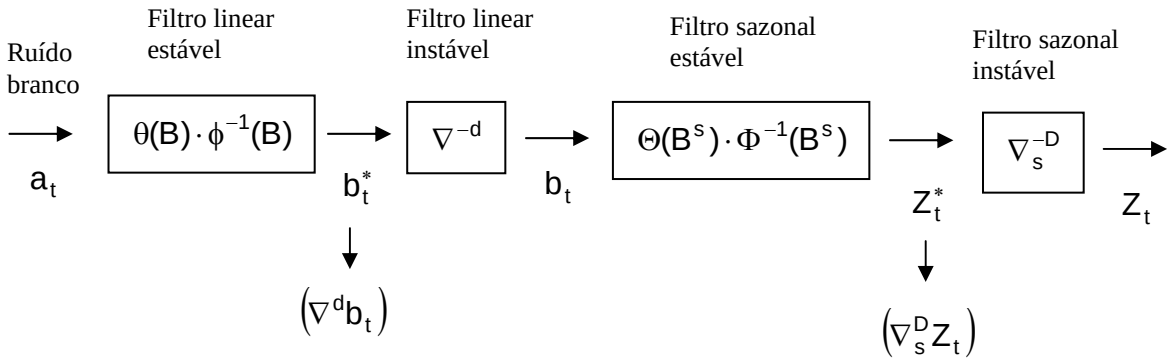

Figura 1 – Representação do modelo ARIMA (p, d, q) x $(P, D, Q)_s$
Fonte: Souza e Camargo (2004:151)

Numa série temporal pode-se encontrar os três filtros ou um subconjunto deles, resultando daí, vários modelos na metodologia Box & Jenkins.
Uma condição que tem que ser colocada no processo estocástico é que este tem que ser estacionário. Um processo estocástico é dito estacionário de segunda ordem quando as seguintes condições forem satisfeitas para qualquer instante de tempo t:

$$E[z_t] = E[z_{t+k}] = \mu\ ,$$
$$Var[z_t] = E[(z_t - \mu)^2] = \sigma^2$$
$$Cov[z_t, z_{t+k}] = E[(z_t - \mu)\cdot(z_{t+k} - \mu)]$$

As duas primeiras condições indicam que a média e a variância de Z_t não variam com o tempo e a terceira, indica que as autocovariâncias não dependem do tempo e sim em relação à distância k que separa as observações.
Quando a série recebe a influência de fatores sazonais, outro tipo de correlação passa a ter importância: a correlação entre os instantes de tempo distantes entre si por s ou múltiplos de s, onde s representa o período da sazonalidade.
A tabela a seguir apresenta as propriedades e características para a identificação teórica dos parâmetros p, q, P e Q dos modelos AR(p), MA(q), ARMA(p,q), SAR(P), SMA(Q) e SARMA(P,Q).

	AR(p)	MA(q)	ARMA(p,q)	SAR(P)	SMA(Q)	SARMA(P,Q)
Modelo expresso em termos dos w_t's anteriores	$\phi(B)w_t = a_t$	$\theta^{-1}(B)w_t = a_t$	$\theta^{-1}(B)\phi(B)w_t = a_t$	$\Phi(B^s)w_t = a_t$	$\Theta^{-1}(B^s)w_t = a_t$	$\Theta^{-1}(B^s)\Phi(B^s)w_t = a_t$
Modelo expresso em termos dos a_t's anteriores	$w_t = \phi^{-1}(B)a_t$	$w_t = \theta(B)a_t$	$w_t = \phi^{-1}(B)\theta(B)a_t$	$w_t = \Phi^{-1}(B^s)a_t$	$w_t = \Theta(B^s)a_t$	$w_t = \Phi^{-1}(B^s)\Theta(B^s)a_t$
Função de Autocorrelação ρ_k	Infinita (exponenciais amortecidas e/ou senóide amortecido). Não se anulam bruscamente.	Finita. Anulam-se bruscamente no lag k.	Infinita (exponenciais amortecidas e/ou senóide amortecidos para k > q-p). Não se anulam bruscamente.	Infinita (exponenciais amortecidas e/ou senóide amortecido). Não se anulam bruscamente.	Finita. Anulam-se bruscamente no lag k.	Infinita (exponenciais amortecidas e/ou senóide amortecidos para k > Q - P). Não se anulam bruscamente.
Função de autocorrelação parcial ϕ_{kk}	Finita. Anulam-se bruscamente no lag k.	Infinita (dominada por exponenciais amortecidas e/ou senóide). Não se anulam bruscamente.	Infinita (dominada por exponenciais amortecidas e/ou senóide amortecidos para k > q-p). Não se anulam bruscamente.	Finita. Anulam-se bruscamente no lag k.	Infinita (domonada por exponenciais amortecidas e/ou senóide). Não se anulam bruscamente.	Infinita (exponenciais amortecidas e/ou senóide para k > Q - P). Não se anulam bruscamente.

Tabela 1 – Comportamento teórico dos modelos AR(p), MA(q), ARMA(p,q), SAR(P), SMA(Q) e SARMA(P,Q)
Fonte: Souza e Camargo (2004:68)

3. KNN (K – *NEAREST NEIGHBORS)*

KNN é um classificador onde o aprendizado é feito de forma supervisionada. O conjunto de treinamento é formado por vetores n-dimensionais e cada elemento deste conjunto representa um ponto no espaço n-dimensional.
Para determinar a classe de um elemento que não pertença ao conjunto de treinamento, o classificador KNN procura K elementos do conjunto de treinamento que estejam mais próximos deste desconhecido elemento, ou seja, que tenham a menor distância.
Estes K elementos são chamados de K-vizinhos mais próximos. Verifica-se quais são as classes desses K vizinhos e a classe mais frequente será atribuída à classe do elemento desconhecido.
A métrica mais comum no cálculo de distância entre dois pontos é a distância Euclidiana cuja definição é:

Seja X= (x_1, x_2, ... , x_n) e Y= (y_1, y_2, ... , y_n) dois pontos do $\Re^n$. A distância Euclidiana entre X e Y é dada por

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + ... + (x_n - y_n)^2} .$$

O KNN é um classificador que possui apenas um parâmetro livre que é o número de K-vizinhos que é controlado pelo usuário com o objetivo de obter uma melhor classificação.
Este processo de classificação pode ser computacionalmente exaustivo se considerado um conjunto com muitos dados. Para determinadas aplicações, no entanto, o processo é bem aceitável.

4. SISTEMA PROPOSTO

Na tomada de decisão sobre a estrutura Box & Jenkins, adequada à modelagem de um determinado processo estocástico, é grande a dose de julgamento subjetivo a ser realizado por um especialista de análise de séries temporais.
O sistema especialista proposto para as estruturas Box & Jenkins de modelos sazonais e não sazonais consiste das variáveis de entrada, que são os vetores formados pelos estimadores da autocorrelação (ACF)e da autocorrelação parcial (PACF), determinados da seguinte forma:

- Modelos Box & Jenkins não sazonais

Utilizam-se os estimadores das autocorrelação e autocorrelação parcial nos *lags* 1, 2, 3, 4 e 5. Estes estimadores são gerados através de simulações de séries dos modelos AR(1), AR(2), MA(1), MA(2) e ARMA(1,1) e Ruído branco com 300, 350, 400, 450, 500 e 600 observações para cada um deles. Para cada série gerada, forma-se o vetor das variáveis de entrada $[ACF1\ ACF2\ ACF3\ ACF4\ ACF5\ PACF1\ PACF2\ PACF3\ PACF4\ PACF5]$, denominado vetor dos estimadores. Ao todo foram usados 3600 vetores de entrada para o conjunto de treinamento e 864 vetores para o conjunto de teste.

- Modelos Box & Jenkins sazonais de período 12

Utilizam-se os estimadores das autocorrelação e autocorrelação parcial nos *lags* 12, 24, 36, 48 e 60. Estes estimadores são gerados através de simulações de séries dos modelos SAR(1), SAR(2), SMA(1), SMA(2) e SARMA(1,1) e Ruído branco com 300, 350, 400, 450, 500 e 600 observações para cada um deles. Para cada série gerada, forma-se o vetor das variáveis de entrada $[ACF12\ ACF24\ ACF36\ ACF48\ ACF60\ PACF12\ PACF24\ PACF36\ PACF48\ PACF60]$, denominado vetor dos estimadores. Ao todo foram usados 3600 vetores de entrada para o conjunto de treinamento e 864 vetores para o conjunto de teste.
Os vetores formados pelos estimadores da autocorrelação e da autocorrelação parcial das séries sazonais de período 3, 4 e 6 consistem dos seguintes *lags*: *lags* 3, 6, 9, 12 e 15, *lags* 4, 8, 12, 16 e 20 e *lags* 6, 12, 18, 24 e 30, respectivamente.
As entradas serão os vetores descritos acima e as saídas serão os modelos AR(1), AR(2), MA(1), MA(2), ARMA(1,1) e Ruído branco no caso dos modelos não sazonais e SAR(1), SAR(2), SMA(1), SMA(2), SARMA(1,1) e Ruído branco no caso dos modelos sazonais.

5- RESULTADOS OBTIDOS E CONCLUSÕES

Para comparar desempenhos, usou-se como critério de avaliação o percentual de acerto do classificador KNN e o obtido pelo *software Forecast Pro for Windows* (FPW) - Versão 3.50, que permite identificar os modelos Box & Jenkins em todos os conjuntos de teste propostos neste trabalho.
Na tabela a seguir, tem-se o percentual de acertos do conjunto de teste das séries não sazonais e sazonais utilizando o classificador KNN.

	Não Sazonal	Sazonal período 3	Sazonal período 4	Sazonal período 6	Sazonal período 12	Média
AR(1) ou SAR(1)	82,64	77,08	76,39	78,47	79,86	78,89
AR(2) ou SAR(2)	77,08	75,7	75,7	73,61	71,53	74,72
MA(1) ou SMA(1)	76,39	76,39	74,31	75	75,7	75
MA(2) ou SMA(2)	68,06	72,92	74,31	77,08	70,83	72,64
ARMA(1,1) ou SARMA(1,1)	35,42	42,36	43,75	34,72	32,64	37,78
Ruído Branco	95,14	97,92	90,28	96,53	98,61	95,7
Média	72,45	73,73	72,45	72,57	71,53	72,46

Tabela 2 – Resultados obtidos utilizando o classificador KNN

Na tabela 3, tem-se o percentual de acertos do conjunto de teste das séries não sazonais e sazonais utilizando o *software* FPW.

	Não Sazonal	Sazonal período 3	Sazonal período 4	Sazonal período 6	Sazonal período 12	Média
AR(1) ou SAR(1)	80,56	68,75	65,97	66,67	67,36	69,86
AR(2) ou SAR(2)	76,39	68,75	61,81	66,67	56,95	66,11
MA(1) ou SMA(1)	77,78	80,56	78,47	83,34	77,78	79,59
MA(2) ou SMA(2)	55,56	59,03	54,86	53,47	54,86	55,56
ARMA(1,1) ou SARMA(1,1)	36,12	39,58	38,89	31,95	27,08	34,72
Ruído Branco	90,28	97,23	91,67	90,97	98,61	93,75
Média	69,45	68,98	65,28	65,51	63,77	66,6

Tabela 3 – Resultados obtidos utilizando o *software* FPW

Pode-se observar que o método proposto obteve resultados siginificativamente melhores que os encontrados através do s*oftware* FPW. Nos dois classificadores, o percentual de acerto da estrutura ARMA(1,1) ou SARMA(1,1) não foram significativos e este resultado deve ser analisado à parte, aplicando o teste de sobrefixação.

6.REFERÊNCIAS BIBLIOGRÁFICAS

ANTUNES, Cláudia M., Aula 6 – Aprendizagem Baseada em Instâncias, disponível em http://mega.ist.utl.pt/~ic-apr/documentos/aulas/aula6_knn_em_svm.pdf, acesso em 20/03/2005.

BERRY, Michael J. A., LINOFF, Gordon, *Data Mining Techniques: for Marketing, Sales, and Customer Support*, John Wiley & Sons, Inc , USA, 1997.

BOX, P.E. G., JENKINS, M.G., *Time Series Analysis Forecasting and Control*, Holden - Day Inc., 1976.

ENDERS, Walter, *Applied Econometric Time Series*, John Wiley & Sons, Inc, USA, 1st ed., 1995.

GNECCO, Bruno Barberi et al., Um Sistema de Visualização Imersivo e Interativo de Apoio ao Ensino de Classificação de Imagens, disponível em http://www.di.ufpb.br/liliane/publicacoes/wrv2001-cave-final.pdf, acesso em 25/05/2005.

HAMILTON, James D., *Time Series Analysis*, Princeton University Press, New Jersey, 1994.

HAN, Jiawei, KAMBER, Micheline, *Data Mining: Concepts and Techniques*, Academic Press, USA, 2001.

HART, Peter, DUDA, Richard O., *Pattern Classification*, John Wiley Professio, USA, 2ª ed., 2000.

HARVEY, Andrew C., *Time Series Models*, , MIT Press, Great Britain 2sd ed. , pp. 22/28, 1993.

LANGIE, Leonardo C., LIMA, Vera L. S., Classificação Hierárquica de Documentos Textuais Digitais Usando o Algoritmo Knn, disponível em http://www.nilc.icmc.usp.br/til2003/oral/Langie_Lima_18.pdf, acesso em 20/03/2005.

MACHADO, Maria Augusta S., Identificação das Estruturas Box & Jenkins não Sazonais usando Redes Neurais Nebulosas, Tese de Doutorado, PUC-RJ, 2000.

MORETTIN, Pedro A., TOLOI, Célia M., *Previsão de Séries Temporais*, Atual, 1987.

PYLE, Dorian, *Data Preparation for Data Mining*, Academic Press, USA, 1999.

REYNOLDS, B., STEVENS T., MELLICHAMP R., Smith M. J., *Box-Jenkins Forecast Model Identification*, A.I. Expert June 1995.

SHUMWAY, Robert H., STOFFER, David S., *Time Series Analysis and Its Applications*, Springer, 2000.

SOUZA, C.R., CAMARGO, M.E., *Análise e Previsão de Séries Temporais: os Modelos ARIMA*, s/e, 2ª edição, Rio de Janeiro, 2004.

WEBB, Andrew R., *Statistical Pattern Recognition*, John Wiley & Sons Ltd, 2º edition, UK, 2002.