APLICAÇÃO DE REDES NEURAIS ARTIFICIAIS NA ANÁLISE DE PADRÕES DE VARIABILIDADE DE MESOESCALA

RODRIGO ANDRADE DE BEM [1]
SILVIA SILVA DA COSTA BOTELHO [1]
MAURÍCIO MAGALHÃES MATA [1]

[1] Fundação Universidade Federal do Rio Grande - FURG
Avenida Itália Km 8, CEP 96201-090 - Rio Grande – RS, Brasil
debem@ecomp.furg.br, silviacb@ee.furg.br, mauricio.mata@furg.br

Abstract. In this paper we apply a Neural Network (NN) to treat large oceanographic datasets, specifically to study the mesoscale variability of an oceanic boundary current. The main objective is to distill the massive oceanographic datasets down to a new space of smaller dimension, characterizing the essential information contained in the data. Due to the natural nonlinearity of those data, traditional multivariate analysis, like the Principal Component Analysis (PCA), may not represent reality. However, Nonlinear Principal Component Analysis (NLPCA) can be performed by a neural network model. This work presents the methodology associated with the use of a multi-layer NN with a bottleneck to extract nonlinear information of the data. We illustrate its good performance with a set of tests against comparisons using this methodology and classical PCA in the Sea Surface Temperature (SST) satellite images of the southwestern Pacific Ocean.

Keywords: analysis methods, NLPCA, neural networks.

1. Introdução

Um típico problema existente na oceonografia é reduzir as dimensões de grandes séries de dados com o objetivo de avaliar melhor o volume de informações contidas nas mesmas. Uma aproximação clássica para resolver esse problema é a utilização de métodos estatísticos lineares como a Análise dos Componentes Principais (PCA). O PCA encontra os autovalores e autovetores da matriz de covariança dos dados e, com esse resultado, pode-se realizar a redução dimensional dos dados e analisar os padrões principais de variabilidade presentes.

O fato do PCA resolver o problema através de uma aproximação linear pode levar a uma simplificação indesejada da variabilidade existente. Essa limitação usualmente é mais significativa em fenômenos oceanográficos com escalas espaciais de centenas de quilômetros e escalas temporais de meses, chamados de fenômenos de mesoescala. Como esses fenômenos afetam fortemente o ecossistema marinho (e assim a distribuição de peixes) uma técnica capaz de levar em consideração a componente não linear dos mesmos é desejável.

A utilização de RNAs tem sido proposta por diversos autores como uma ferramenta para tentar superar as limitações impostas pelo PCA nos problemas oceanográficos Hsieh (2000). Assim, pretende-se avaliar o potencial das RNAs na investigação de padrões de variabilidade em áreas do oceano dominadas por fortes fenômenos de mesoescala. Os resultados são comparados com os obtidos através do método PCA, ressaltando as vantagens e desvantagens da implementação de RNAs.

2. Os Dados

Os dados usados são uma série de três anos e meio (de 1991 até 1994) de imagens de satélite, da Temperatura Superficial do Mar (TSM), do sudoeste do Oceano Pacífico. Estas imagens foram derivadas de imagens com resolução de 1 km x 1 km, gravadas pelo "Advanced Very High Resolution Radiometer" (AVHRR) através dos satélites de órbita polar do "National Oceanic Atmospheric Administration" (NOAA). A precisão radiométrica dessas imagens é em torno de 0.5 °C e a resolução temporal é usualmente 4 imagens por dia.

Como este conjunto de dados é fortemente contaminado por nuvens, ele foi interpolado obtendo-se um grid 9 km x 9 km para cada 10 dias. Este procedimento permite a remoção de nuvens, e ao mesmo tempo, mantém a maior parte das características da TSM que são importantes para este estudo. Além disso, a suposição de que resoluções espacial e temporal, de 9 km x 9 km e 10 dias, respectivamente, são suficientes para o estudo de padrões de mesoescala é válida, e comumente utilizada na oceanografia.

O conjunto de dados interpolados usado no presente estudo tem 128 pontos de temporais e dimensão de 60 x 60 pixels, cobrindo a área desde 31° até 36°S e de 150,5° até 155,5°E, como visto na **Figura 1**. A região foi escolhida por ser uma das mais energéticas dos oceanos do mundo, com um sinal de mesoescala limpo e distinto.

Devido à magnitude dos dados de entrada, uma normalização apropriada foi usada. De cada variável de entrada é subtraída a sua média e dividida pelo seu desvio padrão do valor mais alto da entrada, Hsieh (2000).

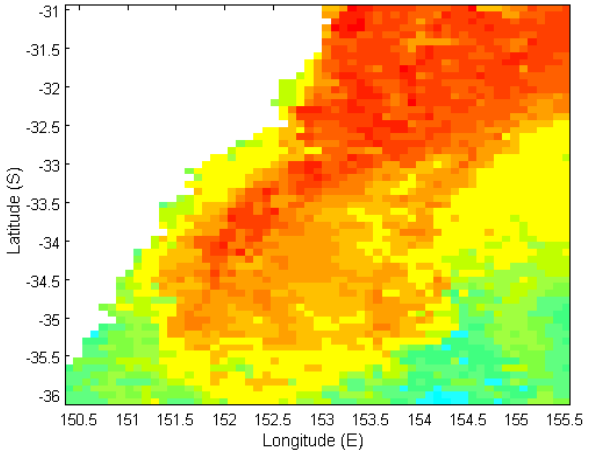

Figura 1: Área de estudo no sudoeste do Oceano Pacífico. A imagem acima é uma das 128 imagens do conjunto de dados analisado. Notam-se várias características de mesoescala (vórtices e meandros) ao longo da linha frontal entre as águas quentes tropicais (vermelho) e as águas mais frias com origem sub antártica (azul).

2. A Teoria do NLPCA

Oceanógrafos adotam geralmente métodos estatísticos tradicionais que têm o objetivo principal é realizar uma redução dimensional em um conjunto bruto de dados oceanográficos, extraindo a informação essencial contida nos mesmos. Tais métodos trabalham encontrando os modos principais de variabilidade em um conjunto das variáveis x, assim reduzindo a dimensionalidade e permitindo a extração da característica de x.

Um destes métodos de extração de características, a Análise dos Componentes Principais (PCA) conhecido também como método das Funções Empíricas Ortogonais (EOF), tem sido usado extensamente nas Ciências de Terra desde sua introdução por Lorenz (1963). O PCA é usado para detectar e caracterizar a estrutura linear ótima, de dimensões reduzidas, de uma série de dados. Entretanto, o PCA e os métodos relacionados podem produzir uma simplificação da série de dados que está sendo analisada devido à suposição que os fenômenos lineares são dominantes. O advento das Redes Neurais Artificiais (RNAs) cria a possibilidade da limitação linear na análise de dados ambientais ser finalmente eliminada.

2.1. Problema da Extração de Características

Uma imagem de satélite pode ser vista como um vetor. Se a largura e a altura da imagem forem w e h pixels respectivamente, o número de elementos (pixels) deste vetor será $w*h$. A construção deste que é chamado de *vetor imagem*, é executada por uma concatenação simples - as linhas da imagem são colocadas uma ao lado outra, como mostrado na **Figura 2**.

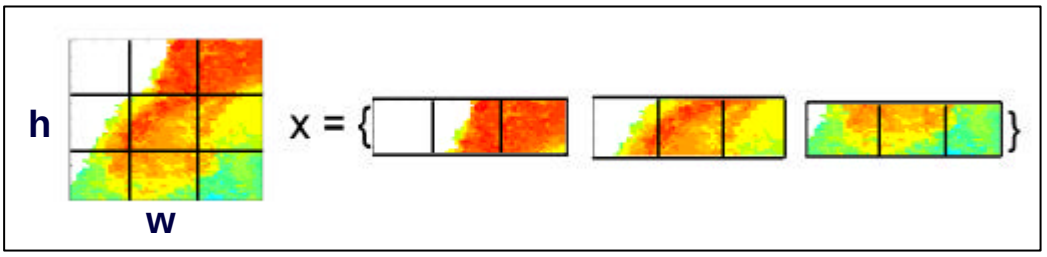

Figura 2: construção do *vetor imagem.*

2.2. Espaço da Imagem

O *vetor imagem* pertence a um espaço, chamado *espaço da imagem*, que é o espaço de todas as imagens cuja dimensão é $w*h$ pixels. Tipicamente nossas imagens de satélite podem ser divididas por diversas características físicas que são muito similares de imagem para imagem (ex.: porções de terra, nuvens, zonas iguais da temperatura, etc.). Assim, ao traçar os *vetores imagem* eles tendem a agruparem-se para formarem a um aglomerado estreito no *espaço da imagem*, que é comum entre eles. Isto é mostrado na **Figura 3** (veja mais detalhes na subseção seguinte).

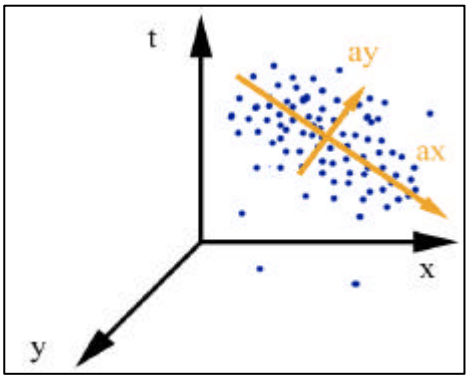

Figura 3: *espaço das imagens* com eixos principais ortogonais (*ax* e *ay*).

2.3. Análise dos Componentes Principais (PCA)

Devido à característica do aglomerado que contém nossos *vetores imagem*, o *espaço da imagem* não é um espaço ótimo para a descrição da imagem de satélite; assim é necessário construir um espaço novo que descreva melhor as imagens. Os vetores base deste espaço novo são chamados de componentes principais. Naturalmente, usar cada pixel pode trazer informações redundantes, porque cada pixel depende de seus vizinhos. Assim, se a dimensão do *espaço da imagem* for $w*h$, então a dimensão do espaço novo é menor do que a dimensão do *espaço da imagem* original. Dessa forma, o objetivo da Análise das Componentes Principais é reduzir a dimensão do conjunto original ou distribuí-lo de modo que a nova base descreva melhor os "modelos" existentes nos dados. Em nosso caso o "modelo" é um

conjunto de padrões de variabilidade de mesoescala. PCA tenta representar a variação total no conjunto das imagens, e explicar esta variação por poucas variáveis. Os vetores novos da base (eixos) serão construídos por uma combinação linear (assim são essencialmente ortogonais). Os componentes nesta base nova não serão correlacionados. Os eixos são mostrados na **Figura 3**. Nós podemos ver que a variação dos dados é máxima na direção ax e assim definido como o primeiro componente principal de variabilidade. O segundo sentido que provoca a variação a maior dos dados, contanto que é ortogonal a ax, é ay.

2.4. Teoria do PCA

Seja $x(t) = [x_1,...,x_p]$ um conjunto de imagens de satélite da superfície do mar, onde cada variável $x_i, (i = 1,..., p)$ é uma série temporal que contém n observações. O PCA é dado por uma combinação linear do x_i, da função temporal u, e de um vetor associado a:

$$u(t) = a * x(t), \qquad \textbf{(1)}$$

então

$$<< \|x(t) - au(t)\|^2 >>, \qquad \textbf{(2)}$$

é minimizado $(<< ... >>)$. Aqui u, chamado de componente principal (PC), é uma série temporal, enquanto a, é o primeiro autovetor da matriz do covariança dos dados, freqüentemente descrever um padrão espacial. Do resíduo, $x - au$, o segundo modo do PCA pode ser obtido, e assim por diante para modos mais elevados.

2.5. A Análise Não Linear

PCA permite somente o mapeamento linear de x para u. Por outro lado, o NLPC é obtido usando uma rede neural multicamadas, veja a **Figura 4**, Kirby e Sirovich (1990). Para executar NLPCA, a RNA tem 3 camadas escondidas de neurônios entre a entrada e a saída. A rede NLPCA de cinco camadas tem p nós na camada da entrada, r nós na terceira camada (do gargalo), e p nós na camada da saída. Os nós na camada 2 e 4 devem ter funções não-lineares da ativação de modo que as camadas 1, 2 e 3 e 3, 4, e 5 possam representar funções suaves arbitrárias. A rede NLPCA permite a compressão de dados porque entradas de dimensão p devem passar pela a camada do gargalo de dimensão r antes de reproduzir as saídas. Uma vez que a rede foi treinada, os valores da ativação do nó do gargalo fornecem os resultados.

Seja $f : \Re^p \to \Re^r$ a função modelada pelas camadas 1, 2 e 3, e seja $s : \Re^r \to \Re^p$ a função modelada pelas camadas 3, 4 e 5. Usando esta notação, os pesos na rede NLPCA são determinados sob a seguinte função objetiva:

$$\min \sum_{l=1}^{n} \|x_l - x_l'\| \qquad \textbf{(3)}$$

, onde x' é a saída da rede. A relação **Equação (1)** é agora generalizada para $u = f(x)$, onde f pode ser qualquer função não-linear explicada por um RNA retro-alimentada da camada da entrada à camada do gargalo, e em vez de **Equação (2)**, $\|x_l - x_l'\|$ é minimizado pelas funções não-lineares de mapeamento, $x' = s(u)$. Do resíduo, $\|x_l - x_l'\|$, o segundo modo do PCA pode ser obtido, e assim por diante para modos mais elevados.

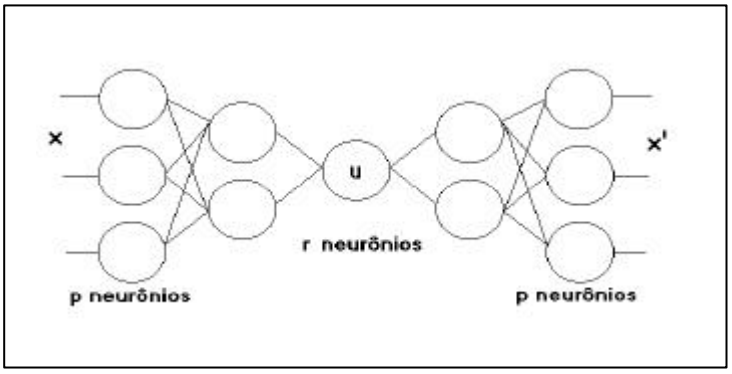

Figura 4: rede NLPCA.

3. Implementação do NLPCA

Os neurônios escondidos têm a função tangente hiperbólica como função não linear de transferência. Para outros os neurônios (entrada, saída e neurônio u) é uma função identidade. A escolha de m, o número dos neurônios escondidos em ambas as camadas escondidas, segue o princípio geral da parcimônia. Aumentos de m tornam maiores a potencialidade modelando não-linear da rede, mas poderiam também conduzir soluções erradas. Usa-se um algoritmo quasi-Newton para realizar a otimização não linear.

4. A Técnica de "Cascateamento"

Os dados apresentam (60x60) 3600 variáveis espaciais e 128 pontos no tempo. Assim, a dimensão dos dados a serem tratados inviabiliza a análise direta através do uso de NLPCA. Com os 3600 neurônios de entrada, a rede NLPCA conteria um número muito grande de parâmetros (pesos associados aos neurônios) em relação à quantidade de amostras temporais (ver Hsieh (2001) para mais detalhes). Reduzir o número de variáveis da entrada, pré-filtrando os dados de TSM utilizando o método PCA foi a solução adotada. Os 3 primeiros PCs (séries temporais) computados da série de dados são mostrados na **Figura 5**.

Olhando esta figura, se pode perceber o amplo domínio do primeiro modo de variabilidade sobre os modos mais elevados. Tendo uma freqüência de aproximadamente 1 ano, o primeiro modo tem um significado físico claro na oceanografia, o qual é relacionado como o aquecimento e resfriamento sazonal das águas de superfície do mar seguindo o ciclo anual de radiação solar. Os segundo e terceiro modos não são passíveis de uma interpretação tão direta.

Esta pré-filtragem, embora necessária, pode estar introduzindo uma simplificação na análise antes mesmo da aplicação do método NLPCA. Assim, com a pretensão de possibilitar a análise direta de grandes áreas de uma maneira confiável, desenvolveu-se a técnica que foi chamada de "cascateamento". Adotando essa metodologia, em nenhum momento da análise dos dados o PCA é utilizado, ou seja, a pré-filtragem não é mais executada. Ao invés disso, as variáveis de entrada foram divididas em pequenos grupos, que são usados diretamente como entrada do NLPCA, respeitando-se assim as limitações da rede. Os modos dessa forma obtidos são agrupados e por sua vez usados como variáveis de entrada de novas redes NLPCA. Esse processo é repetido até que se tenha um único modo, ao final, representando dados de entrada originais.

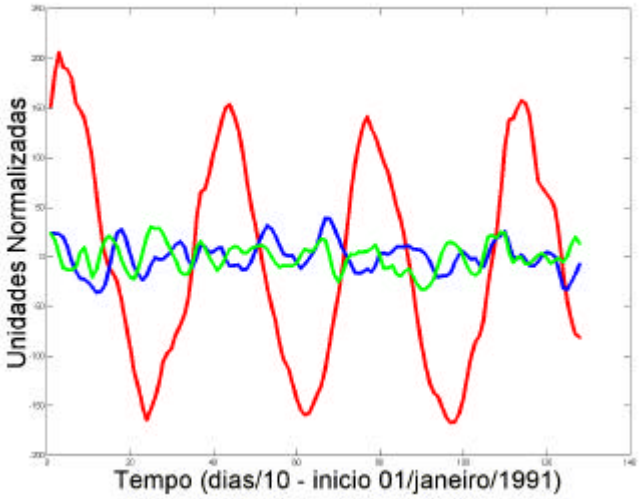

Figura 5: primeiros 3 PCs do PCA: PC1 (vermelho), PC2(azul) e PC3 (verde).

variáveis de entrada foram divididas em pequenos grupos, que são usados diretamente como entrada do NLPCA, respeitando-se assim as limitações da rede. Os modos dessa forma obtidos são agrupados e por sua vez usados como variáveis de entrada de novas redes NLPCA. Esse processo é repetido até que se tenha um único modo, ao final, representando dados de entrada originais.

Apesar de estarem sendo analisados separadamente no primeiro momento, a relação de vizinhança entre as variáveis de entrada, sempre é levada em consideração no passo seguinte, já que os resultados são sucessivamente agrupados. A metodologia é ilustrada na **Figura 6**.

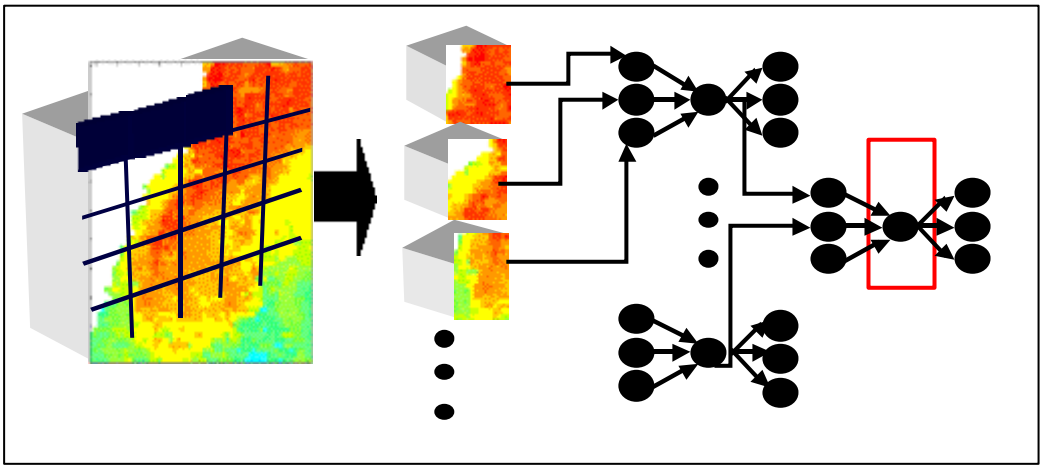

Figura 6: técnica de "cascateamento". O neurônio destacado em vermelho fornece o modo correspondente aos dados de entrada.

5. Resultados e Discussão

Como a solução do PCA para o primeiro modo é linear, e o mesmo domina inquestionavelmente a variabilidade na série de dados (o teste padrão sazonal da oscilação da temperatura de superfície do mar), se esperaria que o primeiro componente principal da rede de NLPCA, tanto com pré-filtragem, quanto através do "cascateamento", fossem similares. Essa expectativa é confirmada na **Figura 7**, onde fica clara a semelhança entre os NLPC1s obtidos com as duas técnicas, e o PC1, já mostrado na **Figura 5**.

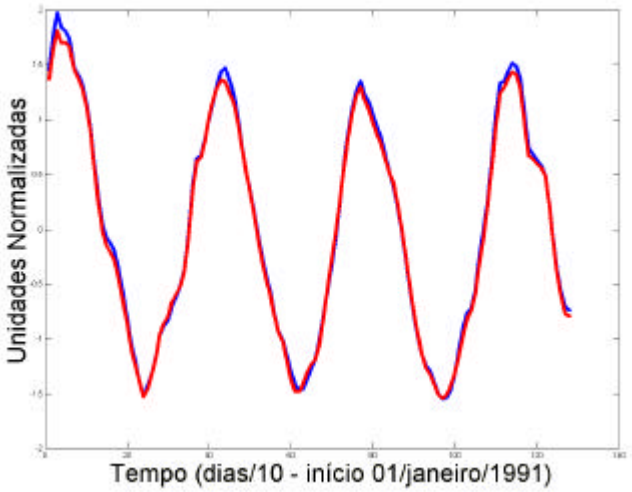

Figura 7: NLPC1 com pré-filtragem (azul) e NLPC1 pela técnica do "cascateamento" (pontos vermelhos).

Inversamente, se pode esperar padrões diferentes ao comparar a solução de PCA para as modalidades mais elevadas (modalidade 2 por o exemplo) com os NLPCs correspondentes. Embora tais modos ainda não tenham sido calculados, tem-se a expectativa que a eliminação da etapa de pré-filtragem na análise dos dados proporcionará ganhos ainda maiores do que aqueles já conseguidos com o NLPCA.

5. Conclusões

No estudo atual, investigou-se o desempenho da rede neural proposta por Hsieh (2000) especificamente estudar padrões de variabilidade de mesoescala. Mostrou-se na seção 3 que os métodos estatísticos acima mencionados, como o PCA, conduzem à aproximação boa da variabilidade somente para o modo dominante, porque é basicamente linear e representa o aquecimento e resfriamento anual da camada de superfície do oceano. As modalidades mais elevadas, entretanto, têm uma natureza não-linear e, em conseqüência, o PCA não pode inteiramente isolar essas modalidades umas das outras e a computação conduz à série temporal que contém mais de um sinal associado a cada processo físico distinto.

Com o uso do NLPCA, as não linearidades existentes nos dados são melhor representadas, o que constitui um ganho na análise. Tal ganho, como mostrado na seção 4, ainda pode estar sendo limitado pelo fato de que se faz necessário uma etapa de pré-filtragem antes da execução do NLPCA sobre uma quantidade grande de variáveis. Assim, a técnica de "cascateamento" mostra-se promissora, já que possibilita a análise de grandes regiões (grande número de variáveis de entrada), sem que se faça uso da análise linear (PCA) em nenhum momento. Como se pode ver na **Figura 7**, os primeiros resultados evidenciam que a técnica é pertinente, já que o modo encontrado está de acordo com o esperado.

As próximas metas para o trabalho são, a obtenção de modos mais elevados com o uso da técnica de "cascateamento" e comparação desses com modos já obtidos com a utilização das outras técnicas. Caso seja comprovado um ganho significativo na análise, a técnica de "cascateamento" estará se mostrando ainda mais eficaz do que a aplicação simples do método NLPCA.

Referências

Hsieh, W. Nonlinear canonical correlation analysis by neural network. Neural Networks, 13:1095-1105, 2000.

Hiesh, W. Nonlinear principal componente analysis by neural neteworks. Tellus, 53A:599-615, 2001.

Kirby, M.; Sirovich, L. Application of karhunen-loeve procedure for the caracterization of humam faces. IEEE transactions on pattern analysis and machine intelligence, 12, 1990.

Lorenz, E. N. Deterministic nonperiodic flow. Atmos. Sci., 20:130-141, 1963.