Seleção de Arquiteturas Neurais Autoassociativas pela Afinidade entre Matrizes

David Pinto*, André P. Lemos†, Antônio P. Braga†
*Programa de Pós-Graduação em Engenharia Elétrica †Departamento de Engenharia Eletrônica
Universidade Federal de Minas Gerais
Belo Horizonte, MG, Brasil
Email: davidem1990@ufmg.br, {andrepl, apbraga}@cpdee.ufmg.br

***Resumo*— Os métodos tradicionais de seleção de complexidade aplicados à análise das componentes principais com redes neurais autoassociativas, como o percentual de variabilidade retida e a validação-cruzada, avaliam as arquiteturas candidatas segundo critérios de erro de representação. A principal deficiência dessa abordagem está na exigência de conjuntos de dados de teste. Caso a comparação seja estabelecida a partir dos próprios padrões de treinamento, as arquiteturas sobre-ajustadas serão favorecidas e a rede escolhida não será genérica. Muitas vezes, no entanto, a quantidade de padrões selecionados para teste pode ser insuficiente na inferência do melhor modelo de extração de características. Neste trabalho é introduzida uma nova abordagem de seleção de complexidade que não requer conjuntos de teste. O método proposto baseia-se na retenção estrutural dos dados no espaço das escores. Diferentemente das técnicas existentes, o método avalia diretamente o grau de redução dimensional da rede e independe da saída gerada.**

I. Introdução

A análise das componentes principais (PCA[1]) [1, 2, 3], formulada inicialmente por Pearson [4], se estabeleceu nas últimas décadas como o principal método multivariado para análise de dados, com uma vasta gama de aplicações. Desde sua reestruturação ao estágio de utilização atual efetuada por Hotelling [5] em 1933, PCA vem sendo empregado com êxito na solução de diversos problemas, tais como monitoramento de processos [6], controle de qualidade [7], análise exploratória de dados [8, 9], detecção, isolamento e reconstrução de sensores falhos [10, 11], visualização de dados [12], recuperação de valores perdidos [13], e até mesmo como ferramenta de visão computacional para compressão de imagens [14] e reconhecimento facial [15].

O principal motivo do sucesso da técnica de PCA está no seu poder de redução dimensional. PCA faz uma projeção dos dados em um novo espaço com número de graus de liberdade reduzido, porém preservando a estrutura de correlação entre as dimensões e capturando de forma ótima a variabilidade presente nos padrões. Dessa forma, o método fornece uma representação simplificada que aprimora a compreensão da estrutura característica dos dados. No entanto, por se restringirem ao mapeamento de correlações lineares entre as dimensões, as técnicas de análise das componentes principais não se adequam satisfatoriamente à solução de problemas não-lineares [16], muito frequentes em quase todas as disciplinas, química, biologia, engenharia, meteorologia e demais. Xu et al. [17] demonstra que, quando PCA é aplicado na extração de características de conjuntos não-linearmente correlacionados, as componentes menos expressivas (ou seja, aquelas que teoricamente poderiam ser suprimidas da análise) nem sempre se associam a ruído ou variância desprezível, muito pelo contrário, podem preservar conteúdo estrutural de relevância equivalente à das componentes principais. No entanto, se por um lado o descarte dessas dimensões resulta em perda de informação, mantê-las indica que PCA necessita de muitas componentes para se tornar plausível na solução do problema.

As limitações do método linear motivaram o desenvolvimento de técnicas generalizadoras do PCA padrão, dando origem à metodologia de análise das componentes principais não-lineares (NLPCA[2]). Essa nova abordagem utiliza a mesma formulação empregada na análise linear tradicional, exceto pelo fato de representar os dados por meio de curvas suaves unidimensionais, determinadas através das relações não-lineares entre as dimensões. Tais curvas têm por objetivo minimizar o erro de projeção, isto é, os desvios ortogonais em relação aos dados e maximizar a representação da variabilidade. Dentre as proposições de maior destaque na literatura estão o método de curvas principais [18], a análise via redes neurais autoassociativas [19], PCA orientado a funções de *Kernel* [20] e PCA probabilístico [21].

O método baseado em redes neurais artificiais, proposto por Kramer [19], opera NLPCA através do treinamento de um *perceptron* multicamadas autoassociativo para realizar um mapeamento de identidade dos dados, onde os alvos da saída são as próprias entradas. A presença de uma camada interna de dimensão reduzida força a representação compacta dos padrões apresentados à rede, porém conservando sua estrutura. NLPCA funciona então como um algoritmo de propósito geral para extração de características, capaz de reter o máximo de informação do conjunto de dados original, para determinado grau de compressão. Na visão de Kramer, a principal diferença entre PCA e NLPCA está no fato de que o último envolve mapeamentos não-lineares entre o espaço original e o projetado. Caso existam correlações não-lineares entre as variáveis, NLPCA descreverá os dados com maior acurácia e/ou utilizando para isso um número menor de componentes que PCA.

A arquitetura autoassociativa proposta por Kramer, referida por Kambhatla e Leen [22] como uma técnica não-linear global de redução dimensional, dispõe de uma camada de nós sigmoidais para mapear os dados de entrada e traduzir

[1]Principal Component Analysis

[2]Nonlinear Principal Component Analysis

o problema, seja qual for a sua complexidade, para um contexto tratável linearmente. Kramer [19] sugere a utilização de critérios de informação de Akaike [23] para definir a dimensão dessa camada não-linear, de forma a alcançar capacidade representacional adequada e evitar *overfitting*. No entanto, um quesito essencial no âmbito da análise das componentes principais permanece em aberto. Tendo definido satisfatoriamente a complexidade do mapeamento, é preciso estabelecer o grau reducional mais cabível à solução do problema, ou seja, o número de neurônios na camada de compressão. Monahan [24] propõe o aumento gradual desse número até atingir uma fração preestabelecida da variância total dos dados na saída da rede. Harkat et al. [25] reformula o método da variabilidade total não-reconstruída, desenvolvido originalmente por Qin e Dunia [26] para a análise linear, a fim de selecionar a arquitetura ótima em termos da reconstrução das dimensões dos dados. Scholz [27] infere a complexidade ótima avaliando as arquiteturas candidatas segundo a capacidade em predizer valores faltantes nos dados. O bom desempenho de grande parte dos métodos tradicionais de seleção depende, no entanto, da utilização de conjuntos extra de teste, e, consequentemente, que estes detenham uma quantidade suficiente de padrões para a representação adequada do problema.

O presente trabalho visa, portanto, introduzir uma nova maneira de inferir a dimensionalidade ideal do espaço projetado, independente de conjuntos de teste. O método aqui proposto parte do princípio de que a matriz projetada pela camada de compressão da rede, caso o grau de redução dimensional e o número de neurônios de mapeamento seja compatível com a complexidade do problema, conserva a estrutura da matriz dos padrões de entrada. Tal retenção é representada na forma da matriz de afinidades [28]. Os resultados aqui discutidos revelam que, ao atingir o nível de compressão dimensional adequado, as matrizes de afinidades das entradas e das projeções se alinham. O número de dimensões a ser escolhido corresponderá então ao ponto a partir do qual a curva que relaciona o número de neurônios da camada de compressão à medida desse alinhamento ultrapassa um valor de corte preestabelecido.

O restante do artigo encontra-se organizado da seguinte forma. Na Seção 2, o método de análise das componentes principais lineares é revisado segundo a abordagem minimizadora do erro de projeção. Em seguida, é apresentada a formulação análoga da técnica não-linear baseada em redes neurais. A Seção 3 traz o conceito de matriz de afinidades juntamente com a descrição da métrica para cálculo de alinhamento. Além disso, é apresentado o método de seleção baseado na similaridade entre matrizes. A Seção 4 exibe os resultados obtidos para a técnica de seleção proposta perante a aplicação sobre três bases de dados de classificação. Em adição, é estabelecida uma comparação de desempenho entre o reconhecimento de padrões usando a base original e a base projetada pela camada de compressão da arquitetura escolhida. Por fim, as discussões e conclusões são apresentadas na Seção 5.

II. Abordagem Não-Linear

A. Análise das Componentes Principais

Considerando uma matriz de dados $X_{n\times m} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m]$, com n observações e m variáveis, PCA consiste em uma fatoração ótima de X em outras três matrizes, a matriz de *scores* $T_{n\times f}$ e a matriz de *loadings* $P_{m\times f}$, mais uma de resíduos $R_{n\times m}$, da seguinte forma:

$$X = TP^T + R = \widehat{X} + R \tag{1}$$

onde f corresponde ao número de fatores mantidos para representação da variabilidade nos dados, e as matrizes $\widehat{X}$ e R representam, respectivamente, as parcelas modelada e não-modelada de X.

É comum, de acordo com Kramer [19], visualizar PCA, tomando-se $P^TP = I$ sem perda de generalidade, como um mapeamento linear dos dados de $\mathbb{R}^n$ para $\mathbb{R}^f$ da forma:

$$\mathbf{t} = \mathbf{x}P \tag{2}$$

onde $\mathbf{x}$ representa uma linha (amostra/observação) de X e $\mathbf{t}$ a linha correspondente de T, ou seja, as coordenadas de $\mathbf{x}$ no plano das componentes principais. Os elementos de P, *loadings*, atuam como os coeficientes das transformações lineares. A perda de informação resultante do mapeamento pode ser inferida através da reconstrução do vetor de medidas pela reversão da projeção de volta ao domínio $\mathbb{R}^n$:

$$\widehat{\mathbf{x}} = \mathbf{t}P^T \tag{3}$$

onde $\widehat{\mathbf{x}} = \mathbf{x} - \mathbf{r}$ é o vetor de medidas reconstruído. Quanto menor o número de dimensões do espaço característico, maior será o erro resultante medido através da norma Euclidiana da matriz residual, $\|R\|$.

B. Arquitetura Autoassociativa e NLPCA

No desenvolvimento de NLPCA, o mapeamento para o espaço característico é generalizado, para permitir funcionais não-lineares arbitrários. Analogamente a (2), busca-se então uma representação da forma:

$$\mathbf{t} = \mathbf{g}(\mathbf{x}) \tag{4}$$

onde $\mathbf{g}$ é um vetor composto por f funções não-lineares individuais, $\mathbf{g} = [g_1, g_2, \ldots, g_f]$, análogas às colunas da matriz de *loadings*, P. Dessa forma, considerando t_i o i-ésimo elemento de $\mathbf{t}$, tem-se

$$t_i = g_i(\mathbf{x}). \tag{5}$$

Em comparação com o caso linear, g_1 é referido como o fator não-linear primário, e g_i como a i-ésima componente não-linear de $\mathbf{x}$.

A transformação inversa, reestabelecendo a dimensionalidade original dos dados, analogamente a (3), é implementada por um segundo vetor de funções não-lineares $\mathbf{h} = [h_1, h_2, \ldots, h_m]$:

$$\widehat{x}_j = h_j(\mathbf{t}). \tag{6}$$

A perda de informação é novamente medida por $R = X - \widehat{X}$ e, semelhantemente ao método de PCA, as funções $\mathbf{g}$ e $\mathbf{h}$ são selecionadas visando a minimização de $\|R\|$.

De acordo com Cybenko [29], é possível aproximar qualquer função contínua não-linear $\mathbf{v} = f(\mathbf{u})$, para um grau arbitrário de precisão, por meio de funções da forma:

$$v_k = \sum_{j=1}^{N_2} w_{jk2}\sigma\left(\sum_{i=1}^{N_1} w_{ij1}u_i + \theta_{j1}\right) \tag{7}$$

onde $\sigma(x)$ é qualquer função monotonicamente crescente com $\sigma(x) \to 1$ quando $x \to +\infty$ e $\sigma(x) \to 0$ quando $x \to -\infty$, como por exemplo uma sigmóide:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{8}$$

Unindo (7) e (8), obtém-se a descrição de uma rede neural artificial *feedforward* com N_1 entradas, uma camada intermediária composta por N_2 neurônios com função de transferência sigmoidal do tipo tangente hiperbólica, e um nó de saída linear, o qual calcula, para cada k, a soma de suas entradas. O termo θ corresponde ao viés nodal, tratado como termo ajustável, assim como os pesos sinápticos. A arquitetura com uma camada de nós sigmoidais e duas camadas de conexões ponderadas é suficiente para alcançar a propriedade de aproximação universal.

Basicamente, a capacidade da rede neural em aproximar funções não-lineares arbitrárias depende da presença de uma camada intermediária com nós não-lineares. Na ausência da mesma, ou com a utilização de nós lineares, a rede é capaz apenas de produzir combinações lineares das entradas, dado que a camada de saída também é linear. Partindo dessa consideração, Kramer [19] representa **g** e **h** de acordo com a arquitetura de 5 camadas apresentada na *Figura 1*, organizada em: camada de entrada, camada de mapeamento, camada de compressão, camada de demapeamento e camada de saída.

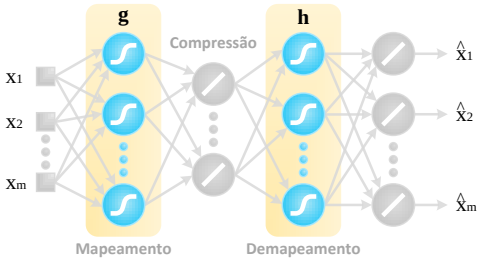


Figura 1: Arquitetura autoassociativa para implementação de NLPCA.

A rede contém 3 camadas escondidas, a de mapeamento, envolvida na modelagem de **g**, a central, cujas saídas representam as características **t**, e a de demapeamento, envolvida na modelagem de **h**. A segunda camada intermediária é designada camada de compressão (*Bottleneck Layer*) por possuir a menor dimensionalidade. A presença de um número menor de neurônios força a rede a desenvolver uma representação compacta dos dados de entrada. Funções de transferência sigmoidais não são exigidas nesta camada, a não ser que se deseje uma resposta mais limitada no espaço característico. Já nas outras duas camadas escondidas, de mapeamento e demapeamento, os nós devem, necessariamente, possuir funções de transferência não-lineares, para modelagem das funções **g** e **h**.

III. Afinidade entre Matrizes

Uma matriz de afinidades deve representar as relações entre elementos e grupos de elementos de um conjunto de dados. Dado um conjunto de dados $X = \{\mathbf{x}_i\}_{i=1}^{N}$, onde N é o número de amostras, não necessariamente rotuladas, o elemento a_{ij} da matriz de afinidades $A = [a_{ij}]_{i,j=1}^{N}$ contém a **afinidade** do par $(\mathbf{x}_i, \mathbf{x}_j)$. Para afinidades reflexivas, a matriz A é simétrica, ou seja, $a_{ij} = a_{ji}$.

As afinidades são tipicamente representadas por alguma medida quantitativa de distância, como por exemplo a Distância Euclidiana, descrita conforme (9). Assim, os elementos da matriz podem ser representados na forma $A(i,j) = d_E(\mathbf{x}_i, \mathbf{x}_j)$.

$$d_E(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{(x_{11}, x_{21})^2 + \cdots + (x_{1n} - x_{2n})^2} \tag{9}$$

A construção da matriz de afinidades é ilustrada usando o conjunto de dados clássico de Fisher [30]. A base, denominada *Iris*, consiste de três classes, *Virginica*, *Versicolor* e *Setosa*, cada uma contendo 4 características e 50 observações. Cada elemento a_{ij} da matriz de afinidades mede a distância entre os vetores da i-ésima e j-ésima linhas do conjunto de dados. Logo, quanto menor o valor de a_{ij}, maior é a proximidade entre as amostras e, consequentemente, maior é a similaridade entre o par $(\mathbf{x}_i, \mathbf{x}_j)$. A Figura 2 mostra as matrizes de afinidades para os padrões originais e para os padrões projetados pela camada de compressão da rede neural autoassociativa. Veja que os elementos da diagonal principal são todos iguais a zero, já que a distância entre um vetor e ele mesmo é nula. Ao utilizar três neurônios na camada escondida central, percebe-se uma grande similaridade entre as matrizes de afinidades, mostrando que a rede foi capaz de reter a estrutura dos dados de entrada no espaço das escores, mesmo suprimindo uma dimensão.

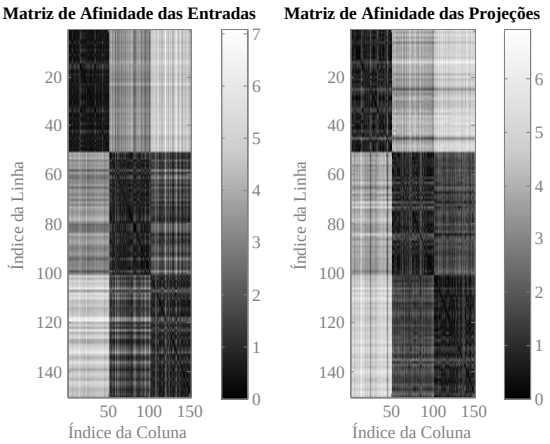


Figura 2: Matriz de Similaridade pela distância Euclidiana para a base de dados Iris [30] utilizando 3 neurônios de compressão.

A. Alinhamento entre Matrizes de Afinidades

Uma forma de quantificar a similaridade entre duas matrizes de afinidades consiste em calcular o Alinhamento Empírico entre ambas. Tal métrica, proposta por Cristianini et al. [31], é definida como segue:

$$A(I,P) = \frac{\langle I, P \rangle_F}{\sqrt{\langle I, I \rangle_F \langle P, P \rangle_F}} \tag{10}$$

onde I e P são as matrizes de afinidades das entradas e das projeções, respectivamente, e $\langle .,. \rangle$ corresponde ao produto interno de Frobenius [31],

$$\langle I, P \rangle_F = \sum_{i=1}^{m} \sum_{j=1}^{m} I(i,j) P(i,j) \tag{11}$$

Em geral, quanto maior a similaridade entre as duas matrizes analisadas, maior é a proximidade do critério de alinhamento frente ao valor unitário. Dessa forma, medidas superiores a $0,9$ são um bom indicativo da retenção estrutural dos dados ao comparar a matriz de afinidades dos padrões de entrada à matriz de afinidades dos padrões projetados pela camada central da rede autoassociativa.

B. Seleção das Componentes pelo Alinhamento entre Matrizes de Afinidades

A rede neural autoassociativa pode ser vista como uma técnica de análise das componentes principais porque consegue extrair a estrutura característica dos dados, separando-a da parcela ruidosa existente, com o mínimo de perda de informação. No entanto, para garantir uma boa representatividade dos dados, com erro de projeção mínimo, é preciso estabelecer um grau de redução dimensional adequado. Com as funções de mapeamento e demapeamento já devidamente ajustadas, sabe-se que, a partir de um determinado número de neurônios de compressão, o qual se espera ser menor que a dimensionalidade dos dados (e geralmente é o que ocorre), a aproximação gerada na saída da rede acompanha satisfatoriamente os dados de entrada. Entretanto, a grande dificuldade é determinar a quantidade de nós que marca exatamente a extração apenas das componentes geradoras dos dados, isto é, aquelas associadas à variabilidade pertinente dos padrões.

O que determina na verdade a boa qualidade representacional da rede é a sua capacidade de conservar a estrutura dos dados de entrada no decorrer das camadas. Dessa forma, quando a dimensão dos nós de compressão permitir uma representação satisfatória, a retenção estrutural ficará nítida. Uma forma de avaliá-la consiste em comparar a matriz de entrada X com a matriz das projeções T. No presente trabalho tal comparação é estabelecida através do cálculo do alinhamento entre as matrizes de afinidades de ambas. Quanto maior a similaridade entre tais matrizes, maior a retenção estrutural.

O método de seleção proposto consiste, portanto, em variar gradualmente o número de neurônios da camada de compressão e quantificar a similaridade entre as matrizes de afinidades das entradas e das respectivas projeções. Quando o grau de compressão adequado for atingido, o valor do alinhamento ficará próximo de 1, mantendo-se nesse nível para todas as demais quantidades de neurônios. O gráfico relacionando o número de nós de compressão ao alinhamento consistirá então de uma curva com um ponto que cruza o valor de corte preestabelecido para a métrica de similaridade entre as matrizes. Tal ponto indica diretamente a quantidade ideal de neurônios para a extração da estrutura característica dos dados. O procedimento de implementação do método se encontra detalhado a seguir:

Passos do Método:

- **Passo 1:** Selecionar e normalizar (entre 0 e 1) a base de dados alvo;
- **Passo 2:** Treinar a rede completa (arquitetura da Figura 1), usando algum algoritmo de otimização como o *backpropagation* [32], para ajustar as funções de mapeamento e demapeamento. Nesta etapa é importante estabelecer a dimensão adequada das camadas não-lineares. A seleção a partir de critérios de informação estatísticos de Akaike, como efetuado por Kramer [19], fornece resultados satisfatórios.
- **Passo 3:** Fragmentar a rede treinada e descartar a parte responsável pela expansão dos dados projetados, restando apenas a arquitetura ilustrada na Figura 3;
- **Passo 4:** Operar a nova rede utilizando os mesmos dados de treinamento, para os quais as funções não-lineares foram otimizadas;
- **Passo 5:** Construir a matriz de afinidades dos padrões de entrada, X, e das projeções, T;
- **Passo 6:** Calcular o alinhamento entre as duas matrizes de afinidades, I e P;
- **Passo 7:** Repetir os passos anteriores variando o número de neurônios da camada de compressão de 1 até a dimensionalidade dos dados. Apenas o procedimento de seleção do número de neurônios sigmoidais não precisa ser repetido.

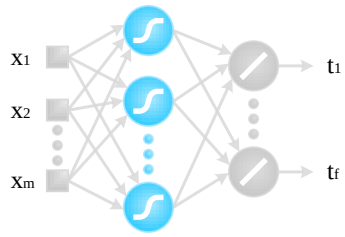


Figura 3: Projeção das entradas no espaço das escores.

Nota-se que o critério utilizado para comparar o desempenho das arquiteturas candidatas não requer conjuntos de validação e teste. A conservação estrutural dos dados depende do quanto de informação a camada de compressão é capaz de propagar, porém, essencialmente como uma medida de sua dimensão, e não do ajuste de seus pesos sinápticos. Isso significa que o método não favorece arquiteturas sobre-ajustadas, mesmo que a comparação entre as candidatas seja estabelecida usando o próprio conjunto de treinamento. Na verdade, o ideal é ajustar primeiramente as funções de mapeamento e demapeamento à complexidade do problema e depois verificar qual o grau de retenção provido pela camada central da rede. Por isso, o mais indicado é utilizar a própria massa de dados aplicada no treinamento das camadas não-lineares da rede.

IV. Resultados

Os métodos tradicionais de seleção das componentes principais da técnica de NLPCA com redes autoassociativas, como aqueles baseados no percentual de variabilidade retida e na validação-cruzada, avaliam as arquiteturas candidatas segundo critérios que contrastam a saída da rede ao alvo estabelecido para tal. A principal deficiência dessa abordagem está na exigência de conjuntos de dados de teste. Caso a comparação seja estabelecida a partir dos próprios padrões de treinamento, as arquiteturas sobre-ajustadas serão favorecidas e a rede escolhida não será genérica no tratamento do problema. Muitas vezes, no entanto, a quantidade de padrões selecionados para teste pode ser insuficiente na inferência do modelo ótimo em

termos da extração das características principais do problema em questão. O método proposto neste trabalho, discutido na seção anterior, não requer conjuntos de teste.

Para avaliar a eficácia da nova abordagem proposta, foram selecionadas 3 bases de dados de classificação do repositório UCI [33]: a base *Cancer* [34], a base *Parkinson* [35] e a base *Heart*. Tais bases contêm apenas duas classes, uma referente ao diagnóstico positivo da doença e outra referente ao diagnóstico negativo. A primeira base avalia os pacientes segundo 30 características, compondo assim um conjunto de dados com 30 dimensões. Já a segunda, contém medidas de 22 características dos supostos enfermos, enquanto a terceira, com a menor dimensionalidade, avalia 13 características. Cada uma das bases foi submetida à técnica de NLPCA variando gradualmente o número de neurônios da camada de compressão, como sugerido na Seção 3. O ajuste de complexidade das camadas não-lineares foi realizado de acordo com o critério de Akaike descrito em [19]. Quanto ao treinamento das redes, foi utilizado o algoritmo de aprendizado em lote denominado *resilient propagation* (Rprop) [36]. Como critérios de parada foram estabelecidos 5.000 épocas de treinamento ou erro de representação nulo, além da aplicação do mecanismo de parada antecipada [37]. Os resultados obtidos são apresentados na Figura 4. Para cada dimensão do espaço projetado, ela ilustra a mediana e os percentis de 25% e 75%, assim como os valores mínimos e máximos, para um total de 50 experimentos.

Nota-se que, a partir de determinado número de nós de compressão, menor que a dimensão da base, a curva de alinhamento se estabiliza, mostrando que foi atingido o grau máximo de retenção estrutural dos dados. Tal comportamento indica que é possível conservar as características principais da entrada utilizando uma dimensionalidade reduzida, na forma da matriz das projeções, T. A fim de comprovar então que tal matriz preserva a essência do problema para o grau reducional indicado pelo método de seleção proposto, os padrões contidos em X (entrada da rede autoassociativa) e em T (saída da camada de compressão) foram submetidos a um mesmo classificador, para as 3 bases analisadas.

A dimensão adequada para a matriz T foi estabelecida como o número de neurônios que marca o ponto a partir do qual a curva de alinhamento se mantém acima de $0,95$. Os resultados da classificação das três bases a partir de X (base original) e T (base reduzida), medidos pelos critérios de percentual total de acerto e área abaixo da curva ROC, são exibidos na Tabela I. Um classificador idêntico foi capaz de discernir entre as duas classes das bases reduzidas com média e desvio padrão equiparáveis à da classificação das bases originais. A equivalência dos resultados confirmou, portanto, a escolha de arquiteturas capazes de reter satisfatoriamente a estrutura característica dos dados de entrada.

V. Conclusões e Trabalhos Futuros

O indicativo da eficácia do método proposto não está apenas na estabilização da curva de alinhamento médio. A avaliação do comportamento das arquiteturas segundo o grau de retenção estrutural consiste também em uma medida da robustez reducional das redes. Ao mesmo tempo em que o aumento gradual dos nós de compressão eleva e estabiliza a média da similaridade entre os padrões de entrada e os padrões projetados, a variabilidade intrínseca das arquiteturas vai diminuindo. Os diagramas de caixa da Figura 4 mostram perfeitamente isso.

Ao atingir um alinhamento de $0,95$ os modelos passam a se comportar em torno de uma média e com uma variância muito semelhantes. O método garante, portanto, a escolha da arquitetura mais eficiente em termos da extração das características do problema, já que indica o grau de redução dimensional mais robusto e adequado à retenção estrutural dos dados.

A seleção da complexidade fundamentada na análise de matrizes de afinidades, além de se consolidar como uma técnica robusta na inferência do grau reducional apropriado às arquiteturas autoassociativas, totalmente independente de conjuntos de teste, é um método genérico. Os resultados aqui apresentados avaliaram a aplicação sobre problemas de classificação. No entanto, a validade se estende a qualquer problema que envolva a extração de características baseada no mapeamento de identidade, tais como aqueles enumerados na Seção 1 para a técnica de análise das componentes principais.

Esforços futuros estão focados em compreender melhor o comportamento da curva de alinhamento através da utilização de outras métricas e de mais bases de dados. Além disso, experimentos comparativos devem ser realizados para avaliar a eficácia do método proposto em relação às técnicas tradicionais que utilizam conjuntos de teste.

Agradecimentos

O presente trabalho foi realizado com o apoio financeiro da CAPES - Brasil e do Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq.

Referências

[1] J. Jackson, *A User's Guide to Principal Components*, ser. Wiley series in probability and mathematical statistics: Applied probability and statistics. John Wiley & Sons, 1991.

[2] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed. Springer, oct 2002.

[3] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 2, pp. 37–52, 1987.

[4] K. Pearson, "On lines and planes of closest fit to systems of points in space," *Philosophical Magazine*, vol. 2, no. 6, pp. 559–572, 1901.

[5] H. Hotelling, "Analysis of complex statistical variables into principal components," *Journal of Educational Psychology*, vol. 24, no. 6, pp. 417–441, sep 1933.

[6] M. Piovoso, K. Kosanovich, and J. Yuk, "Process data chemometrics," *Instrumentation and Measurement, IEEE Transactions on*, vol. 41, no. 2, pp. 262–268, 1992.

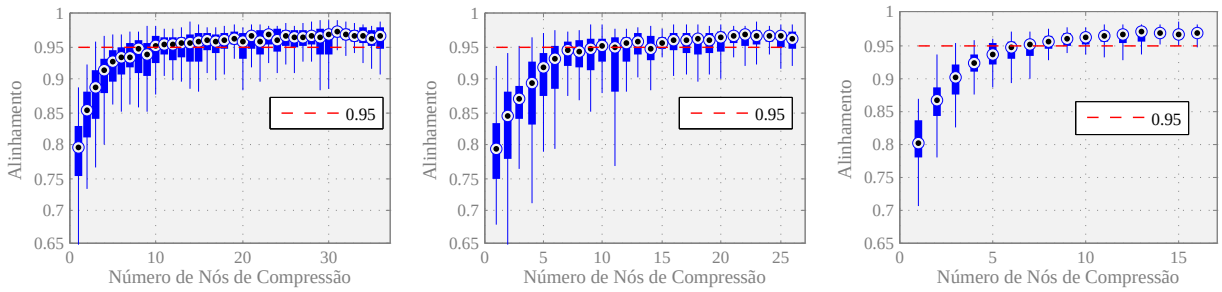
[7] J. F. MacGregor and T. Kourti, "Statistical process control of multivariate processes," *Control Engineering Practice*, vol. 3, no. 3, pp. 403–414, Mar. 1995.

[8] J. W. Tukey, "A data analyst's comments on a variety of points and issues," in *Event-Related Brain Potentials in Man*, E. Callaway, Ed. Academic Press, 1978, pp. 139–151.

[9] J. Himberg, J. Mantyjarvi, and P. Korpipaa, "Using pca and ica for exploratory data analysis in situation awareness," in *Multisensor Fusion and Integration for Intelligent Systems, 2001. MFI 2001. International Conference on*, 2001, pp. 127–131.

[10] B. M. Wise, N. L. Ricker, and D. J. Veltkamp, "Upset and sensor failure detection in multivariate processes," *AIChE Meeting*, nov 1989.

[11] R. Dunia, S. J. Qin, T. F. Edgar, and T. J. McAvoy, "Identification of faulty sensors using principal component analysis," *AIChE Journal*, vol. 42, no. 10, pp. 2797–2812, 1996.



(a) Base *Cancer* (b) Base *Parkinson* (c) Base *Heart*

Figura 4: Curva de alinhamento para as bases de dados analisadas. O valor de corte para seleção da dimensionalidade adequada é de 0.95.

Tabela I: Resultados da classificação das três bases de dados para uma rotina de 50 execuções. Critérios avaliados a partir do conjunto de dados de teste.

Base de Dados	Dimensão Original	Dimensão Reduzida	Acurácia para a Base Original (%)	Acurácia para a Base Reduzida (%)	AUC para a Base Original	AUC para a Base Reduzida
Cancer	30	11	$96,58 \pm 1,32$	$95,52 \pm 1,37$	$0,989 \pm 0,010$	$0,975 \pm 0,017$
Parkinson	22	15	$83,89 \pm 4,24$	$82,91 \pm 3,90$	$0,897 \pm 0,035$	$0,894 \pm 0,034$
Heart	13	7	$80,14 \pm 3,85$	$81,19 \pm 3,71$	$0,802 \pm 0,063$	$0,817 \pm 0,057$

[12] E. Oja, *Subspace methods of pattern recognition*. Research Studies Press, 1983.

[13] H. Martens and T. Naes, *Multivariate Calibration*, ser. Wiley series in probability and mathematical statistics. Wiley, 1992.

[14] C. Clausen and H. Wechsler, "Color image compression using PCA and backpropagation learning," *Pattern Recognition*, vol. 33, no. 9, pp. 1555–1560, 2000.

[15] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, jan 1991.

[16] M. Palus and I. Dvoek, "Singular-value decomposition in attractor reconstruction: Pitfalls and precautions," *Physica D: Nonlinear Phenomena*, vol. 55, pp. 221–234, 1992.

[17] L. Xu, E. Oja, and C. Y. Suen, "Modified hebbian learning for curve and surface fitting," *Neural Networks*, vol. 5, no. 3, pp. 441–457, 1992.

[18] T. Hastie and W. Stuetzle, "Principal curves," *Journal of the American Statistical Association*, vol. 84, no. 406, pp. 502–516, 1989.

[19] M. A. Kramer, "Nonlinear Principal Component Analysis Using Autoassociative Neural Networks," *AIChE Journal*, vol. 37, no. 2, pp. 233–243, Feb. 1991.

[20] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Comput.*, vol. 10, no. 5, pp. 1299–1319, jul 1998.

[21] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 61, no. 3, pp. 611–622, 1999.

[22] N. Kambhatla and T. K. Leen, "Dimension reduction by local principal component analysis," *Neural Comput.*, vol. 9, no. 7, pp. 1493–1516, oct 1997.

[23] H. Akaike, "A new look at the statistical model identification," *Automatic Control, IEEE Transactions on*, vol. 19, no. 6, pp. 716–723, 1974.

[24] A. H. Monahan, "Nonlinear principal component analysis: Tropical indo?pacific sea surface temperature and sea level pressure," *Journal of Climate*, vol. 14, pp. 219–233, 2001.

[25] M.-F. HARKAT, G. Mourot, J. Ragot *et al.*, "Variable reconstruction using rbf-nlpca for process monitoring," *5th IFAC Symposium on Fault Detection, Supervision and Safety for Technical Processes*, pp. 1239–1244, 2003.

[26] S. Qin and R. Dunia, "Determining the number of principal components for best reconstruction," *Journal of Process Control*, vol. 10, no. 2-3, pp. 245–250, 2000.

[27] M. Scholz, "Validation of nonlinear pca," *Neural Processing Letters*, vol. 36, no. 1, pp. 21–30, 2012.

[28] Y. Weiss, "Segmentation using eigenvectors: a unifying view," vol. 2, 1999, pp. 975–982 vol.2.

[29] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Mathematics of Control, Signals, and Systems (MCSS)*, vol. 2, no. 4, pp. 303–314, Dec. 1989.

[30] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[31] N. Cristianini, J. Shawe-Taylor, and J. Kandola, "On kernel target alignment," in *Proceedings of the Neural Information Processing Systems, NIPS'01*. MIT Press, 2002, pp. 367–373.

[32] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Parallel distributed processing: explorations in the microstructure of cognition, vol. 1," D. E. Rumelhart, J. L. McClelland, and C. PDP Research Group, Eds. Cambridge, MA, USA: MIT Press, 1986, ch. Learning internal representations by error propagation, pp. 318–362.

[33] K. Bache and M. Lichman, "UCI machine learning repository," 2013. [Online]. Available: http://archive.ics.uci.edu/ml

[34] O. L. Mangasarian, W. N. Street, and W. H. Wolberg, "Breast cancer diagnosis and prognosis via linear programming," *Operations Research*, vol. 43, no. 4, pp. 570–577, 1995.

[35] M. A. Little, P. E. McSharry, S. J. Roberts, D. A. Costello, and I. M. Moroz, "Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection," *BioMedical Engineering OnLine*, vol. 6, no. 1, 2007.

[36] M. Riedmiller and H. Braun, "A direct adaptive method for faster backpropagation learning: The rprop algorithm," in *IEEE INTERNATIONAL CONFERENCE ON NEURAL NETWORKS*, 1993, pp. 586–591.

[37] N. Morgan and H. Bourlard, "Advances in neural information processing systems 2," D. S. Touretzky, Ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1990, ch. Generalization and parameter estimation in feedforward nets: some experiments, pp. 630–637.