Classificação de Doenças Intersticiais Pulmonares Difusas através de Tomografia Computadorizada de Alta-Resolução

Isadora Cardoso[1], **Heitor Ramos**[1], **Eliana Almeida**[1]

[1]LACCAN/Instituto de Computação – Universidade Federal de Alagoas (UFAL)
Av. Lourival Melo Mota, s/n – Tabuleiro do Martins – CEP 57072-900
Maceió – AL – Brasil

{icps, heitor, eliana.almeida}@ic.ufal.br

Abstract. *The goal of this work is to help the development of a computer-aided diagnosis of lung diseases. In this first stage we used principal component analysis (PCA), linear discriminant analysis (LDA) and k-nearest neighbors algorithm (KNN) to classify 3252 regions of interest (ROI) of High-resolution computed tomography of the chest into 6 lung patterns. From each ROI we extracted 28 features, which were used to evaluate the performance of two dimensionality reduction techniques (PCA and LDA). We further applied KNN ($K = 5$) to classify the ROI into the correspondent lung pattern. We achieved 80,82% correct classification rate using 13 dimensions with PCA and 83,74% using 5 dimensions with LDA.*

Resumo. *O objetivo deste trabalho é auxiliar no desenvolvimento de uma ferramenta de diagnóstico de doenças pulmonares auxiliado por computador. Nessa primeira etapa utilizamos análise de componentes principais (PCA), análise do discriminante linear (LDA) e o algoritmo de k-vizinhos mais próximos (KNN) para classificar 3252 regiões de interesse (ROI) de Tomografias Computadorizadas de Alta-Resolução de tórax em relação à 6 padrões pulmonares. Cada ROI possui um total de 28 dimensões que foram reduzidas por PCA e LDA e então classificadas por KNN (k = 5). Obtivemos uma taxa de classificação correta de 80,82% em 13 dimensões com PCA e 83,74% em 5 dimensões com LDA.*

1. Introdução

Doenças pulmonares intersticiais difusas (DPIDs) são patologias que causam disfunção respiratória. Combinando dados clínicos com dados tomográficos, é possível um alto grau de acurácia no diagnóstico correto. Dentre as técnicas de imagens radiológicas, principalmente na avaliação de DPIDs, destaca-se a Tomografia Computadorizada de Alta-Resolução (TCAR), uma vez que melhora significativamente a sensibilidade e a especificidade do diagnóstico clínico [Elicker et al. 2008], além de reduzir a exposição da estrutura torácica [Pereyra et al. 2014].

De uma base de dados com 3252 regiões de interesse (ROI) obtidas de TCAR do tórax de pacientes, [Pereyra et al. 2014] extraíram 28 atributos relativos à textura para classificar seis padrões pulmonares, a saber: pneumonia (PC), áreas enfisematosas (AE), espessamento de septo (ESI), favo de mel (FM), opacidade em vidro fosco (OVF) e tecido pulmonar saudável (TS). Após, utilizaram o algoritmo k-vizinhos mais próximos (*k-nearest neighbor* - KNN) (k = 5) e obtiveram acurácia média de 80%. Com a mesma

base de dados, [Almeida et al. 2015b] utilizaram o modelo estatístico de mistura de Gaussianas (MMG) e obtiveram uma classificação correta mínima de 60%. Em sequência, [Almeida et al. 2015a] aplicaram MMG nos cinco atributos mais significativos para obter funções de pertinência *fuzzy* e obtiveram uma média de 63% de classificação correta.

Ainda com esse conjunto de dados, o presente projeto tem a intenção de ajudar no desenvolvimento de um sistema de diagnóstico auxiliado por computador (CAD). Nossa primeira etapa é a identificação de técnicas que possam ajudar na redução de dimensionalidade, pois, com todas as 28 dimensões extraídas, é possível ter um *overfitting* e a chamada maldição da dimensionalidade, que gera erros devido à distância em que os pontos se encontram espalhados nas dimensões, gerando uma escassez (pode haver escassez de pontos em algumas partes). Embora reduzindo, é importante atentar para não perder informações importantes presentes nos dados [Tan et al. 2005]. Para isso, utilizamos análise de componentes principais (PCA) e análise do discriminante linear (LDA), por serem duas técnicas já consolidadas e muito utilizadas em vários aspectos de imagens médicas.

2. Materiais e métodos

Foram extraídas de 3252 ROI um conjunto de 28 atributos, a saber: estatísticas de primeira ordem como média, mediana, desvio padrão, assimetria e curtose; 14 atributos de textura de [Haralick et al. 1973] (energia, momento da diferença, correlação, variância, momento da diferença inversa, soma dos quadrados: média, soma dos quadrados: variância, soma dos quadrados: entropia, momento da variância, momento da entropia, medida de informação de correlação 1, medida de informação de correlação 2 e máximo coeficiente de correlação); Medidas de energia de textura de Laws (*Laws' wave measures (rotation-invariant), Laws' ripple measures (rotation-invariant), e Laws' level measures*); medidas estatísticas da densidade spectral de potência (DSP) (média, desvio padrão e mediana); e dimensão fractal. Detalhes desses atributos estão disponíveis em [Rangayyan 2004].

Utilizamos estes atributos e aplicamos as técnicas PCA e LDA. Em sequência, utilizamos um classificador KNN, com k = 5. Para validar, utilizamos o método *hold-dout* de validação cruzada, com cerca de 80% das 3252 ROIs para treino do modelo e cerca de 20% para testes, comparando-a à classificação dos médicos especialistas feitas anteriormente. Em todo o procedimento foi utilizado a linguagem R (versão 3.2.2)[1].

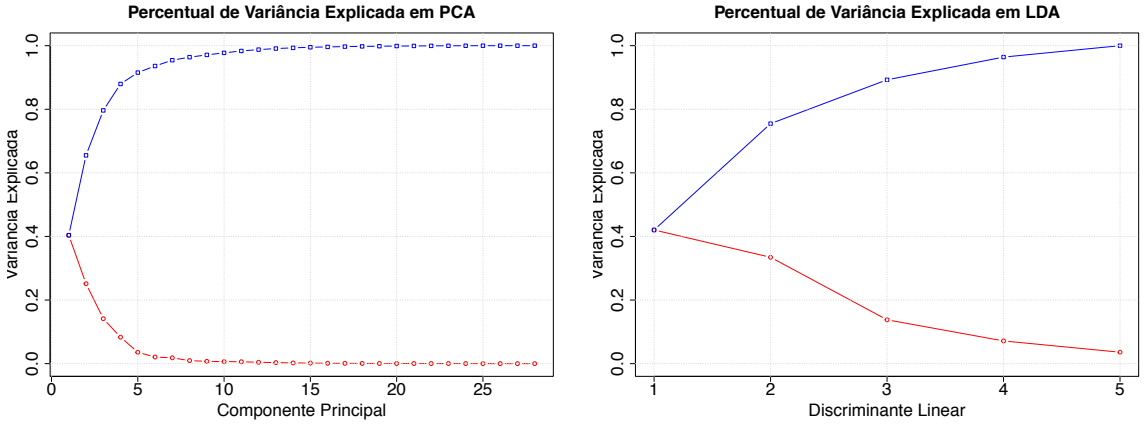
2.1. Análise de Componentes Principais

PCA é uma técnica não supervisionada que, em um conjunto de d dimensões, procura uma transformação de menor dimensão que mantenha o máximo possível da variância dos dados [James et al. 2013].

2.2. Análise do Discriminante Linear

Ao lidar com dimensões paralelas, PCA escolherá uma dimensão que junte todas, o que, claramente, levará a erros de classificação. Nessas situações, LDA é uma melhor escolha. LDA é uma técnica supervisionada que deseja encontrar o hiperplano que maximize a separação entre as médias de cada classe e miniminize a dispersão total das mesmas, o que evita sobreposições que podem piorar a separação [Zaki and Wagner Meira 2014].

[1] http://www.r-project.org



(a) PCA (b) LDA

Figura 1. Quantidade de variância explicada através das 1(a) componentes principais e 1(b) discriminante linear ideais.

2.3. K-vizinhos mais próximos

Um classificador k-vizinhos mais próximos analisa a localização de um ponto no espaço e determina, por seus k vizinhos, qual sua classificação. Isso é feito medindo-se a distância (ex: Euclidiana) do ponto pros seus vizinhos [Tan et al. 2005]. Por simplicidade e para facilitar na comparação das técnicas de redução de dimensionalidade aqui tratadas, utilizamos o KNN para classificar os padrões pulmonares.

3. Resultados

Na figura 1, a linha vermelha representa o quanto de variância cada componente principal e discriminante linear ideais explicam em relação aos dados. A linha azul representa o acumulado de variância que representam ao total. É possível notar pela figura 1 que com cerca de 13 componentes principais já se pode explicar quase 100% da variância dos dados. LDA consegue uma redução ainda maior, 5 dimensões. O lado negativo dessas transformações é que temos que utilizar todo o conjunto de dados para obtê-la.

Na tabela 1 podemos ver a comparação do resultados aplicando as técnicas PCA e LDA e dos dados sem técnica alguma de redução de dimensionalidade, classificados pelo KNN. Na tabela, TP (*true positive*) representa a classificação correta, enquanto FP (*false positive*) representa o erro do tipo I. O Espessamento do septo (ESI) é a doença mais difícil de classificar, obtendo um número mais baixo de TP em todas as técnicas. A classificação dos dados originais e dos obtidos pelas técnicas são próximas, o que mostra quão bom são os conjuntos conseguidos. Em negrito, estão destacados os melhores valores de TP e FP para cada classe. Observe que a técnica LDA apresenta sistematicamente melhores valores quando comparada com PCA ou com os dados originais, ficando pouco abaixo em apenas alguns poucos casos.

4. Conclusão

Neste trabalho utilizou-se PCA e LDA para redução de dimensionalidade. Obtivemos uma taxa de classificação correta de 81,13% com PCA (13 dimensões) e 84,38% com LDA (5 dimensões). Ambas são semelhantes ou superiores à taxa de acerto do conjunto

	Original		PCA		LDA	
	TP	**FP**	**TP**	**FP**	**TP**	**FP**
PC	**0.8686869**	**0.02355072**	0.7979798	0.03558719	0.8484848	0.02669039
AE	0.8111111	0.03014184	0.8111111	0.03019538	**0.8777778**	**0.01950355**
ESI	**0.6796875**	**0.07663551**	0.6718750	0.07865169	**0.6796875**	0.07706767
FM	0.8240741	0.03531599	**0.8333333**	0.03358209	**0.8333333**	**0.03321033**
TS	0.8596491	0.03018868	0.8596491	0.03024575	**0.9122807**	**0.01893939**
OVF	0.8672566	0.02772643	0.8938053	0.02238806	**0.9115044**	**0.01879699**
Total	0.8184109	0.03725986	0.8112923	0.03844169	**0.8438448**	**0.03236805**

Tabela 1. Resultado das classificações do conjunto de dados original, PCA e LDA, com KNN (k = 5)

original, o que cumpre o objetivo. Trabalhos futuros visam utilizar técnicas que explicitem quais atributos são os mais importantes, descartando os que não contribuem tanto. Também visa-se a utilização de mais classificadores a fim de obter melhores resultados.

Referências

Almeida, E., Rangayyan, R. M., and Azevedo-Marques, P. M. (2015a). Fuzzy membership functions for analysis of high-resolution CT images of diffuse pulmonary diseases. In *37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2015, Milan, Italy, August 25-29, 2015*, pages 719–722.

Almeida, E., Rangayyan, R. M., and Azevedo-Marques, P. M. (2015b). Gaussian mixture modeling for statistical analysis of features of high-resolution CT images of diffuse pulmonary diseases. In *2015 IEEE International Symposium on Medical Measurements and Applications, MeMeA 2015, Torino, Italy, May 7-9, 2015*, pages 1–5.

Elicker, B., de Castro Pereira, C. A., Webb, R., and Leslie, K. O. (2008). Padrões tomográficos das doenças intersticiais pulmonares difusas com correlação clínica e patológica. In *Jornal Brasileiro de Pneumologia, Volume 34*, number 9, pages 715–744.

Haralick, R. M., Shanmugam, K., and Dinstein, I. (1973). Textural Features for Image Classification. *Systems, Man and Cybernetics, IEEE Transactions on*, SMC-3(6):610–621.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R*. Springer Texts in Statistics. Springer.

Pereyra, L. C., Rangayyan, R. M., Ponciano-Silva, M., and de Azevedo Marques, P. M. (2014). Fractal analysis for computer-aided diagnosis of diffuse pulmonary diseases in HRCT images. In *2014 IEEE International Symposium on Medical Measurements and Applications, MeMeA 2014, Lisboa, Portugal, June 11-12, 2014*, pages 455–460.

Rangayyan, R. (2004). *Biomedical Image Analysis*. Biomedical Engineering. CRC Press.

Tan, P.-N., Steinbach, M., and Kumar, V. (2005). *Introduction to Data Mining, (First Edition)*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Zaki, M. J. and Wagner Meira, J. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.