1.03.04 - Ciência da Computação / Sistemas de Computação

ESTUDO DE TÉCNICAS DE APRENDIZAGEM DE MÁQUINA PARA ANÁLISE DE TRATAMENTO DE DADOS ORIUNDOS DE IMAGENS MÉDICAS PARA CARACTERIZAÇÃO DA DINÂMICA DAS CIDADES INTELIGENTES

Isadora Cardoso[1], Heitor S. Ramos[2], Eliana Almeida[3]
1. Graduanda em Engenharia de Computação - Instituto de Computação - UFAL
2. Prof. Dr./Orientador - Instituto de Computação - UFAL
3. Profª. Drª. - Instituto de Computação - UFAL

Resumo:

O objetivo deste trabalho é avaliar a viabilidade da construção de um Diagnóstico Auxiliado por Computador para imagens médicas relativas à doenças pulmonares difusas, através de técnicas relativas à aprendizagem de máquina. Utilizou-se técnicas de redução de dimensionalidade e de classificação supervisionada. Conseguiu-se reduzir as 28 dimensões originais para até 5 dimensões. Em termos de diagnóstico, obteve-se acertos de até 86,92%, apresentando resultados superiores aos encontrados na literatura. A diminuição da redundância através da redução de dimensionalidade, a escolha adequada das técnicas de classificação e um treinamento eficaz foram fatores responsáveis pelos resultados superiores aos encontrados na literatura.

Palavras-chave: aprendizagem de máquina, diagnóstico auxiliado por computador, doenças pulmonares difusas

Apoio financeiro: CNPq

Trabalho selecionado para a JNIC pela instituição: UFAL

Introdução:

As cidades estão passando por uma crise socioeconômica devido, principalmente, ao crescimento urbano desordenado e a migração para grandes metrópoles. Isso causa um estresse significativo na estrutura das cidades devido ao aumento da demanda por recursos essenciais, como água, energia, transporte urbano, saúde, educação, segurança, entre outros.

O desafio atual da tecnologia é desenvolver inovações que forneçam tais recursos de forma eficiente a todos. Além disso, a tecnologia deve promover o crescimento social, cultural e sustentável, para aumentar a qualidade de vida dos cidadãos. Dessa forma, são utilizadas diversas Tecnologias da Informação e Comunicação (TICs) para a criação de Cidades Inteligentes (CIs).

Dados relacionados à saúde são de suma importância no contexto de CIs, visto que obter uma saúde de qualidade, e, consequentemente, bons diagnósticos, compõe os direitos do cidadão. Não há dúvidas quanto a importância da utilização de imagens no auxílio a diagnósticos. Essas imagens são uma forma de adquirir informações sobre o paciente. Porém, é comum haver detrimento por ruídos dessas imagens durante sua criação, devido a vários fatores que afetam diretamente seu processo de construção. Como consequência, há dificuldade na aquisição de informações necessárias para o diagnóstico. Técnicas de melhoramento de imagens processadas e avaliadas por computador podem ser utilizadas a fim de ajudar na interpretação do especialista, como um apoio ao diagnósticos -

o Diagnóstico Auxiliado por Computador (CAD).

CADs apresentam diversas vantagens em relação aos seres humanos, como: quando treinados igualmente, computadores não apresentam diferença significativa em suas análises; não são abatidos por sentimentos como cansaço e tédio, por exemplo. Embora vantajosos, CADs não devem substituir o especialista no processo de diagnóstico, uma vez que se fazem presentes fatores subjetivos, como o histórico do paciente, a situação socioeconômica, dentre outros.

Doenças Pulmonares Difusas (DPDs) são um grupo patológico de mais de 150 doenças, que causam disfunção respiratória. DPDs constituem um grande desafio ao pneumologista, devido não só ao grande número de patologias, mas também pela necessidade de conhecimento aprofundado de diversas áreas médicas para fornecer um diagnóstico preciso.

Nesse contexto, esse trabalho se propôs a avaliar a viabilidade da construção de um CAD para imagens médicas relativas à DPDs, avaliando técnicas de aprendizagem de máquina.

Metodologia:

O presente trabalho é derivado de Pereyra *et al* (2014), que utilizaram uma base de dados contendo 3252 regiões de interesse (ROI) obtidas de Tomografia Computadorizada de Alta-Resolução (TCAR) do tórax de pacientes. Dessas imagens, extraiu-se 28 atributos de textura, para classificação referente a seis padrões pulmonares, a saber: (i) pneumonia, (ii) áreas enfisematosas, (iii) espessamento de septo, (iv) favo de mel, (v) opacidade em vidro fosco e (v) tecido pulmonar saudável. Em seguida, os autores aplicaram o algoritmo k-vizinhos mais próximos (KNN) (k = 5) e obtiveram acurácia média de 80%.

Essa mesma base de dados também foi utilizada por Almeida et al (2015), que aplicaram o modelo estatístico de mistura de Gaussianas (GMM) e obtiveram uma classificação correta mínima de 60%. Complementar a este trabalho, Almeida et al (2015) aplicaram GMM nos cinco atributos mais significativos para obter funções de pertinência *fuzzy* e obtiveram uma média de 63% de classificação correta.

Utilizando essa mesma base de dados, nossa primeira etapa foi identificar técnicas que possam ajudar na redução de dimensionalidade, a fim de evitar a maldição da dimensionalidade e reduzir a redundância. Para isso, utilizou-se as seguintes técnicas: (i) análise de componentes principais (PCA), (ii) análise do discriminante linear (LDA) e (iii) técnicas stepwise - *forward* (SSF), (iv) *backward* (SSB) e (v) *forward-backward* (SSFB).

A partir dos subconjuntos obtidos pelas técnicas para redução de dimensionalidade, utilizou-se três algoritmos de classificação, a saber: KNN, GMM e máquina de vetores de suporte (SVM).

Todas a computação foi realizada na linguagem R (versão 3.3.2), devido a sua confiabilidade e precisão ao lidar com funções estatísticas.

A validação dos dados foi feita através de método *holdout* de validação cruzada, com 80% dos dados para treino e 20% dos dados para testes.

Resultados e Discussão:

Na tabela 1, podemos ver quantas dimensões resultaram ao aplicar cada algoritmo e quantas dimensões temos no conjunto de dados original. Os resultados estão em ordem crescente.

LDA apresenta o menor número de dimensões. Porém, LDA combina todos os atributos do conjunto de dados para obter um novo conjunto com informações mais relevantes para o problema. Assim, por se tratar de uma transformação, ainda seria necessário extrair todos os atributos de textura. A viabilidade do processo de extração dos 28 atributos para o CAD proposto não foi estudada neste trabalho.

A tabela 2 apresenta os melhores resultados obtidos pelas técnicas de classificação, juntamente com o conjunto utilizado para obtenção do resultado. TP (*true positive*, verdadeiro positivo) representa a classificação correta, enquanto FP (*false positive*, falso positivo) representa o erro do tipo I.

Tabela 1: Comparação das dimensões resultantes dos algoritmos

Conjunto	Dimensões
LDA	5
SSF	11
SSFB	12
PCA	13
SSB	18
Original	28

Tabela 2: Melhores classificações obtidas

Classificador	Conjunto	TP	FP
KNN	LDA	84,28%	3,26%
SVM	SSB	85,67%	2,92%
GMM	SSFB	86,92%	2,71%

Entre as técnicas utilizadas, GMM apresenta sistematicamente um melhor resultado, embora os resultados obtidos sejam bem próximos no geral. As taxas de falso positivo são abaixo de 3,5% e todos os resultados de correta classificação média são acima de 80%.

Se compararmos os resultados obtidos com os da literatura, nota-se que as taxas de acerto obtidos neste trabalho são superiores, mesmo com erro do tipo I baixo. Tal resultado se deve ao uso das técnicas aplicadas para a redução de dimensionalidade, visto que KNN e GMM foram anteriormente utilizados.

Conclusões:

O presente projeto objetivou uma avaliação da viabilidade da construção de um CAD para imagens médicas relativas a DPDs. Para isso, primeiramente avaliou-se técnicas para redução de dimensionalidade. Dos 28 atributos obtidos originalmente, conseguiu-se uma redução de até 5 dimensões, ao se utilizar LDA. Em seguida, classificou-se os conjunto de dados com KNN, GMM e SVM, que obtiveram um acerto médio superior a 84%. Valores semelhantes, porém sensivelmente inferiores foram obtidos na literatura. Portanto, podemos considerar que nossos resultados foram bem sucedidos.

Em trabalhos futuros será aplicada uma 10-*fold* validação cruzada, a fim de aumentar a confiança dos resultados. Também serão aplicadas mais técnicas tanto para reduzir dimensão, quanto para classificar, a fim de realizar um estudo mais exaustivo sobre o problema, objetivando melhores resultados.

Referências bibliográficas

PEREYRA, L. C.; RANGAYYAN, R. M.; PONCIANO-SILVA,M.; AZEVEDO-MARQUES, P. M., **Fractal analysis for computer-aided diagnosis of diffuse pulmonary diseases in HRCT images**, In: MEDICAL MEASUREMENTS AND APPLICATIONS (MeMeA), 2014, IEEE International Symposium on, p. 1–6.

ALMEIDA, E.; RANGAYYAN, R. M.; AZEVEDO-MARQUES, P. M., **Gaussian mixture modeling for statistical analysis of features of high-resolution CT images of diffuse pulmonary diseases**, In: MEDICAL MEASUREMENTS AND APPLICATIONS (MeMeA), 2015, IEEE International Symposium on, p. 1–5.

RANGAYYAN , Rangaraj M. **Biomedical Image Analysis**, 2004, CRC Press, 1312 p.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introduction to Data Mining**, 1. ed. Boston, 2005, Addison-Wesley. 769 p.