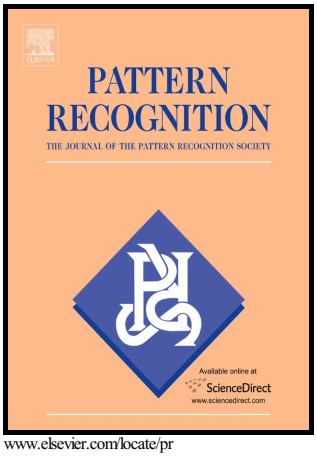
A Multi-kernel based framework for heterogeneous feature selection and over-sampling for computer-aided detection of pulmonary nodules

Peng Cao, Xiaoli liu, Jinzhu Yang, Dazhe Zhao, Wei Li, Min Huang, Osmar Zaiane



www.elsevier.com/locate/pr

PII: S0031-3203(16)30355-7
DOI: http://dx.doi.org/10.1016/j.patcog.2016.11.007
Reference: PR5950

To appear in: *Pattern Recognition*

Received date: 27 June 2016
Revised date: 13 October 2016
Accepted date: 5 November 2016

Cite this article as: Peng Cao, Xiaoli liu, Jinzhu Yang, Dazhe Zhao, Wei Li, Min Huang and Osmar Zaiane, A Multi-kernel based framework for heterogeneous feature selection and over-sampling for computer-aided detection of pulmonary nodules, *Pattern Recognition,* http://dx.doi.org/10.1016/j.patcog.2016.11.007

A Multi-kernel based framework for heterogeneous feature selection and over-sampling for computer-aided detection of pulmonary nodules

Peng Cao[a,b], Xiaoli liu[a,b], Jinzhu Yang[a], Dazhe Zhao[a,b], Wei Li[b], Min Huang[c,d], Osmar Zaiane[e]

[a] *College of Computer Science and Engineering, Northeastern University, Shenyang, China*
[b] *Key Laboratory of Medical Image Computing of Ministry of Education, Northeastern University, Shenyang, China*
[c] *College of Information Science and Engineering, Northeastern University, Shenyang, China*
[d] *State Key Laboratory of Synthetical Automation for Process Industries , Northeastern University, Shenyang, China*
[e] *Computing Science, University of Alberta, Edmonton, Alberta, Canada*

Abstract

Classification plays a critical role in False Positive Reduction (FPR) in lung nodule Computer Aided Detection(CAD). To achieve effective recognition of nodule, many machine learning methods have been proposed. However, multiple heterogeneous feature subsets, high dimensional irrelevant features, as well as imbalanced distribution between the nodule and non-nodule classes typically makes this problem challenging. To solve these challenges, we proposed a multi-kernel based framework for feature selection and imbalanced data learning in Lung nodule CAD, involving multiple kernel learning with a $\ell_{2,1}$ norm regularizer for heterogeneous feature fusion and selection from the feature subset level, a multi-kernel feature selection based on pairwise similarities from the feature level, and a multi-kernel over-sampling for the imbalanced data learning. Experimental results demonstrate the effectiveness of the proposed method in terms of Geometric mean(G-mean) and Area under the ROC curve (AUC), and consistently outperform the competing methods.

Keywords: Lung nodule detection, false positive reduction, classification, imbalanced data learning, multi-kernel learning, feature selection

1. Introduction

Lung cancer is one of the main public health issues in developed countries[1], and early detection of solitary pulmonary nodules (SPNs) is an important clinical indication for early-stage lung cancer diagnosis because SPNs have high probabilities to become malignant nodules [2]. SPNs refer to lung tissue abnormalities that are roughly spherical with round opacity and a diameter of up