Biomedical Signal Processing and Control 40 (2018) 1–9

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Research paper

Lung nodule classification using local kernel regression models with out-of-sample extension

Guohui Wei [a], He Ma [a,b,*], Wei Qian [a,c], Fangfang Han [a], Hongyang Jiang [a], Shouliang Qi [a], Min Qiu [d]

[a] Sino-Dutch Biomedical and Information Engineering School, Northeast University, Shenyang, China
[b] Key Laboratory of Medical Image Computing, Ministry of Education, Shenyang, China
[c] Department of Electrical and Computer Engineering, University of Texas, El Paso, TX, USA
[d] Affiliated Hospital of Jining Medical University, Jining 272029, China

ARTICLE INFO

Article history:
Received 23 November 2016
Received in revised form 6 July 2017
Accepted 24 August 2017

Keywords:
Spectral clustering
Linear regression
Kernel trick
Lung nodules
Out-of-sample

ABSTRACT

Computer-aided classification is a major research task for computer-aided diagnosis of pulmonary nodules. In radiology domain, labeled data can be expensive to generate. Therefore, in this study, a novel unsupervised spectral clustering algorithm was presented to distinguish benign and malignant nodules. In this algorithm, a new Laplacian matrix was constructed by using local kernel regression models (LKRM) and incorporating a regularization term, the regularization term can tackle the out-of-sample problem. To verify the feasibility of our algorithm, a ground truth dataset was assembled from the LIDC-IDRI database, including 371 benign and 375 malignant lung nodules. All nodules were represented by the texture features, which were computed from the regions of interest (ROIs). Extensive experiments on lung nodules showed that the proposed algorithm not only achieved a higher classification performance than existing popular unsupervised algorithms, but also had superiority comparing to some supervised algorithms (linear discriminant analysis and extreme learning machine).

1. Introduction

Lung cancer remains the leading cause of death around the world. Most patients are diagnosed with lung cancer in the late-stage, which cannot be effectively treated. In order to enhance patients' 5-year survival, early diagnosis of lung cancer is very important [1]. Up to now, computed tomography (CT) scan is an effective screening tool for detecting the presence of pulmonary nodules and non-small cell lung cancer [2,3]. However, due to the perceptual error or misinterpretation, it is difficult for radiologists to differentiate malignant nodules from benign ones in CT images. Hence, it is required for a computer-aided diagnosis (CAD) system, which can help to distinguish benign and malignant nodule lesions [4]. Thus, computer-aided classification of benign and malignant tumors, especially the lung nodule images, becomes a popular research domain [5,6].

The major challenges for image classification typically include feature extraction and classification. In feature extraction, current research mainly focuses on designing new features to improve the description and differentiation of images [5]. However, these feature extraction methods cannot solve the intra-class variation and inter-class ambiguity problem. For classification, with the development of medical imaging technology, a number of unlabeled data are generated. However, the current major classifiers need labels to train. Some unsupervised classifiers cannot adapt to the development of data. Therefore, in this paper, we propose a novel unsupervised classification algorithm to classify unlabeled lung nodules.

Several classifiers have been applied to classify benign and malignant lung nodules, such as linear discriminant analysis (LDA) [5,7], artificial neural network [8], and support vector machine (SVM) [9,10]. Lee et al. [7] integrated genetic algorithms with LDA for classifying nodules. Han et al. [9] analyzed 3-D image-based texture features and utilized SVM for nodule diagnosis. Huang and Tu [11] used a logistic regression classifier and open research platform of radiomics for classifying malignancy of small pulmonary nodules. Recently, our group [12] proposed a content-based image retrieval method to measure the similarity of lung masses for classification of benign and malignant lung nodules. Shen et al. [13] introduced a deep learning model of multi-crop convolutional neural networks to tackle the challenging problem of lung nod-

* Corresponding author at: Sino-Dutch Biomedical and Information Engineering School, Hunnan Campus, Northeastern University, Shenyang 110169, China.
E-mail address: mahe@bmie.neu.edu.cn (H. Ma).

http://dx.doi.org/10.1016/j.bspc.2017.08.026
1746-8094/

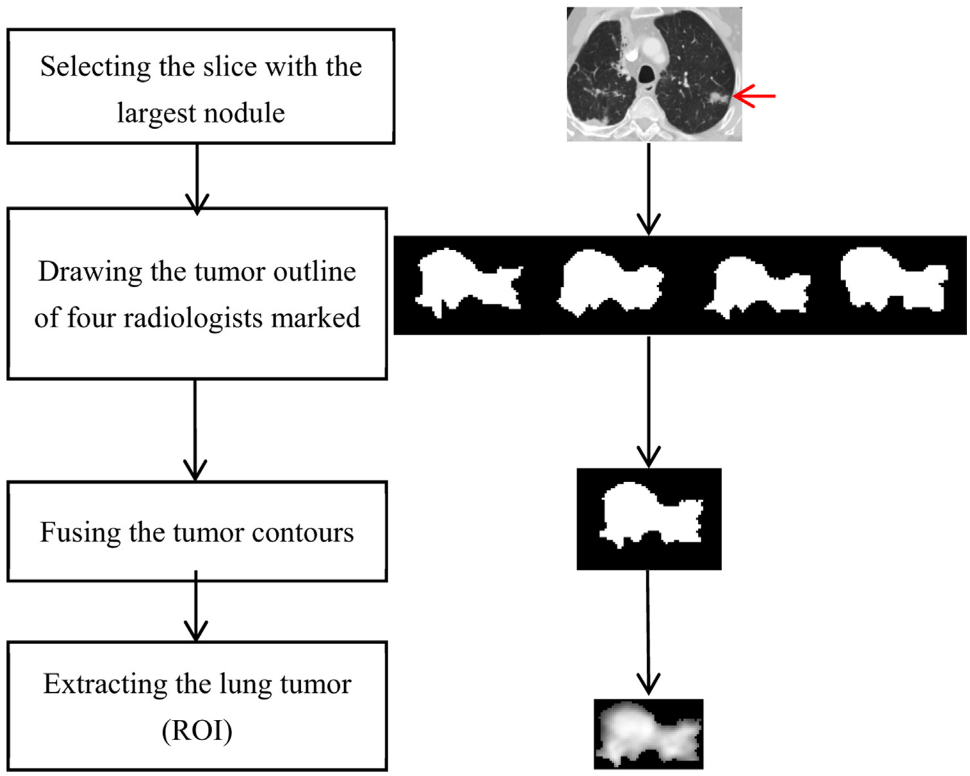

Fig. 1. The workflow of LIDC-IDRI processing.

Table 1
Statistics spectral clustering algorithms' characteristics.

Algorithms	Theoretical basis	Application
NCut	Graph partitioning	Image Segmentation
LLC	Local learning	Clustering
LDMGI	Local discriminant models	Image Clustering
SEC	Local Regression	Image Clustering

ule malignancy suspiciousness classification. Tajbakhsh et al. [14] compared lung nodule detection and classification performance between massive-training artificial neural networks and convolutional neural networks. Liu et al. [15] constructed multi-view convolutional neural networks for lung nodule classification. All these classifiers need data labels to train. However, it is expensive to generate labeled data in radiology domain, which are marked by pathological experiments or radiologists. A number of data have no labels. There are some unsupervised algorithms classifying unlabeled data. Lee et al. [16] proposed an unsupervised learning algorithm of convolutional deep belief networks for learning hierarchical representations from unlabeled images. Zhou et al. [17] used a spectral clustering algorithm to extract live tumors accurately. Han et al. [18] explored vector quantization scheme for accurate segmentation of the lungs and lung nodule detection. However, few unsupervised algorithms are used to classify lung nodules. Jia et al. [19] explored the growth changes feature and a rule-based classifier for benign and malignant lung nodule classification. Hence, we want to construct an unsupervised classifier that is applicable to unlabeled lung nodule data. Spectral clustering is one of the unsupervised classifiers.

Spectral clustering (SC) has gradually become an important clustering technique, which is applicable to many practical cases [20–24]. An overview of the spectral clustering algorithms is provided in Table 1. Normalized cut (NCut), the famous SC algorithm [20], obtained promising clustering performances in many classification applications. Recently, local learning based clustering (LLC) [21] estimated the label of each data point according to its neighbors. The clustering using local discriminant models and global integration (LDMGI) [22] learned a Laplacian matrix by utilizing both local discriminant information and manifold structure. Spectral embedded clustering [23] studied the out-of-sample extension with a linear regression. SC algorithms showed better performance in distinguishing complex data structures compared with traditional clustering algorithms. However, many algorithms, such as Ncut, LLC, LDMGI, suffer from out-of-sample problem that they cannot predict cluster labels to testing data, which is outside training data [23]. This limits the application of spectral clustering in computer-aided classification.

To address the aforementioned issues, in this study, we propose a new SC algorithm which is referred to as local kernel regression models with out-of-sample extension (LKRME) algorithm for lung nodule classification. A new Laplacian matrix is learned from local kernel regression models (LKRM) and global integration to exploit the locally and globally discriminative information. Moreover, a linear regression regularization term (LR) or a global kernel regression regularization term (KR) is integrated into the learned matrix to assign cluster labels to out-of-sample data. Our proposed algorithm differs from the spectral embedded clustering [23], which designs a Laplacian matrix using local regression models and handles the out-of-sample problem with LR, while the LKRME algorithm uses LKRM to construct a new Laplacian matrix, LR and KR to deal with the out-of-sample problem.

The major findings of this study are four aspects:

1) We assemble a ground truth lung nodule dataset from the LIDC-IDRI lung CT image database. Each lung nodule is extracted according to nodule outlines drawing by more than three thoracic radiologists.
2) We propose a novel SC algorithm LKRME to differentiate unlabeled benign and malignant lung nodules. We demonstrate that it is effective to classify unlabeled pulmonary nodules based on clustering algorithm.