Journal of Biomedical Informatics xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

A weighted rule based method for predicting malignancy of pulmonary nodules by nodule characteristics

Aydın Kaya *, Ahmet Burak Can

Hacettepe University, Computer Engineering Department, 06800 Ankara, Turkey

ARTICLE INFO

Article history:
Received 28 December 2014
Revised 18 April 2015
Accepted 15 May 2015
Available online xxxx

Keywords:
Nodule characteristic
Ensemble classifier
Rule based classification
Unbalanced data

ABSTRACT

Predicting malignancy of solitary pulmonary nodules from computer tomography scans is a difficult and important problem in the diagnosis of lung cancer. This paper investigates the contribution of nodule characteristics in the prediction of malignancy. Using data from Lung Image Database Consortium (LIDC) database, we propose a weighted rule based classification approach for predicting malignancy of pulmonary nodules. LIDC database contains CT scans of nodules and information about nodule characteristics evaluated by multiple annotators. In the first step of our method, votes for nodule characteristics are obtained from ensemble classifiers by using image features. In the second step, votes and rules obtained from radiologist evaluations are used by a weighted rule based method to predict malignancy. The rule based method is constructed by using radiologist evaluations on previous cases. Correlations between malignancy and other nodule characteristics and agreement ratio of radiologists are considered in rule evaluation. To handle the unbalanced nature of LIDC, ensemble classifiers and data balancing methods are used. The proposed approach is compared with the classification methods trained on image features. Classification accuracy, specificity and sensitivity of classifiers are measured. The experimental results show that using nodule characteristics for malignancy prediction can improve classification results.

1. Introduction

Lung cancer is one of the leading causes of cancer related deaths worldwide [29]. In the diagnosis of lung cancer, detection of solitary pulmonary nodules is a challenging process for radiologists. A solitary pulmonary nodule is a lung lesion with a diameter of about 2–30 mm and indistinct boundaries. These nodules are generally found fortuitously on tomography scans [1]. With the advances in screening technologies, detection rate of nodules are increased. Computer Aided Diagnosis (CAD) Systems are developed to help radiologists as a second reader. There are two main concerns on CAD Systems; detection and classification of nodules. The challenge in evaluation of a patient's nodule is to determine whether it's benign or malignant. Diagnoses made by radiologists are highly subjective and can be significantly different depending on the level of radiologists' experience.

One of the difficulties in this research area is to find well organized and consistent data. Publicly accessible Lung Image Database Consortium (LIDC) database [2] provides researchers with CT images, nodule region of interests and nodule characteristics as radiographic descriptors. In this database, all cases are evaluated by four radiologists. Each radiologist gives his/her estimations for the boundaries and characteristic ratings of nodules. As Zinovev et al. [9] states, radiologist anonymity and lack of ground truth in LIDC database are challenges; however the database provides the opportunity to build different computer aided diagnosis methods.

In this study, a weighted rule based method for malignancy prediction on pulmonary nodules is presented. The first goal of the study is to show the usefulness of nodule characteristics in malignancy prediction. Separate datasets are defined for each nodule characteristic by applying majority voting on LIDC data. Since most datasets are highly unbalanced, data balancing methods are applied on the datasets. In addition, features are ranked for each nodule characteristic. Subsequent to feature ranking, different feature set sizes are determined for each characteristic by using average ranks and success rate ratios. Since ensemble classifiers are used as a tool to handle unbalanced datasets [31,34,24,25], nodule characteristics are classified with ensemble classifiers. A separate ensemble classifier is built for each nodule characteristic. In the ensemble classification, LDA [35], SVM [28], kNN [37] Adaboost [38], and Random Forest [36] classifiers are tested as the base classifier. Outputs of ensembles are used as inputs for a weighted rule

* Corresponding author. Tel.: +90 312 297 75 00/131; fax: +90 312 297 75 02.
E-mail addresses: aydinkaya@cs.hacettepe.edu.tr, aydinkaya83@gmail.com (A. Kaya), abc@cs.hacettepe.edu.tr (A.B. Can).

http://dx.doi.org/10.1016/j.jbi.2015.05.011
1532-0464/© 2015 Published by Elsevier Inc.

based method, where the rules are constructed from radiologists' evaluations on nodule characteristics for previous cases. Thus, the expert opinion in LIDC dataset is utilized in the rule based method to predict malignancy. The correlation between malignancy and other nodule characteristics are analyzed to understand the importance of each characteristic in malignancy determination. Furthermore, evaluations of radiologists for the same nodule are analyzed to figure out level of agreement among experts in relation to nodule characteristics. The general schema of the proposed work is shown in Fig. 1.

In later sections, we first give some information about related work. Then, in the methodology section, we give details on LIDC dataset, extracted image features, dataset balancing, feature extraction and feature size determination, and classification steps. After presenting experiments, we discuss the results of our work and future plans.

2. Related work

Most studies on lung nodule detection and classification use only image features to classify lung nodules. With databases like LIDC and NELSON Trial, different challenges have emerged. Handling multiple annotator assessments, providing objective evaluation, predicting other nodule characteristics besides malignancy, and using semantic characteristics to improve malignancy prediction are some of these challenges. We give brief information about some studies which use the LIDC dataset and deal with nodule characteristics. Detailed literature information can be found in surveys by Suzuki [3], Sluimer et al. [4], El-Baz et al. [5].

Zhao et al. [6] propose a CAD system for estimating malignancy of nodules. In this system, ensembles of linear classifiers with feature subsets are constructed. Majority voting is applied on classifier outputs to find probability of malignancy. Jabon et al. [8] develop a content and semantic based image retrieval system that takes CT images as input and retrieves similar images by using image features and semantic characteristics. Euclidean and cosine similarity measures are used in this study. Median voting is used to find the summarized rating for a nodule with multiple annotators. Zinovev et al. [7] propose a method which uses an ensemble decision tree classifier with active learning to predict nodule characteristics. Active learning uses radiologists' agreements on characteristic ratings and predicts the nodules on which radiologists do not agree. The results obtained are better compared to

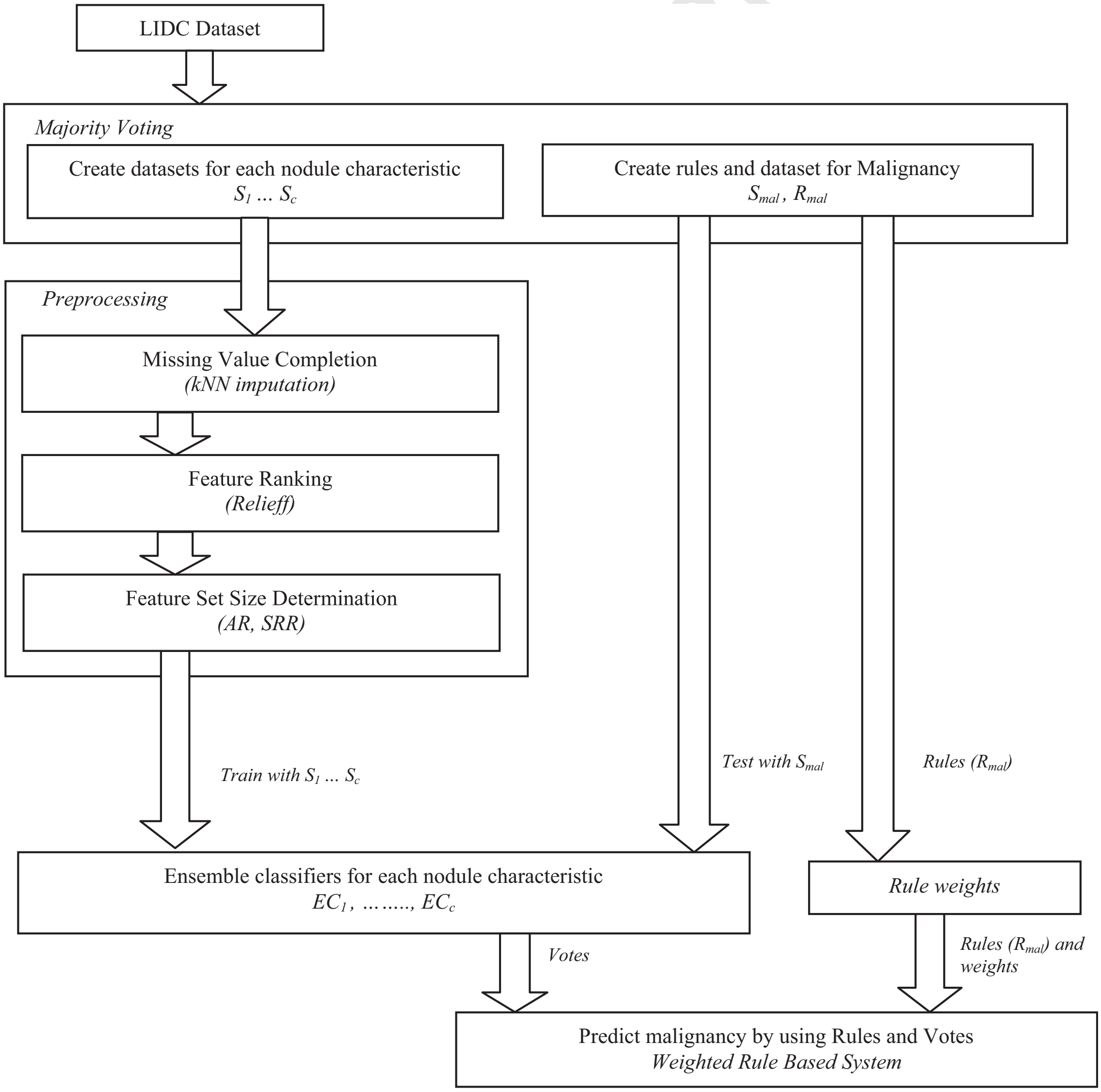
LIDC Dataset
Majority Voting
Create datasets for each nodule characteristic
$S_1 \dots S_c$
Create rules and dataset for Malignancy
S_{mal}, R_{mal}
Preprocessing
Missing Value Completion
(kNN imputation)
Feature Ranking
(Relieff)
Feature Set Size Determination
(AR, SRR)
Train with $S_1 \dots S_c$
Test with S_{mal}
Rules (R_{mal})
Ensemble classifiers for each nodule characteristic
$EC_1, \dots\dots, EC_c$
Rule weights
Votes
Rules (R_{mal}) and weights
Predict malignancy by using Rules and Votes
Weighted Rule Based System

Fig. 1. General schema of the proposed method.

Please cite this article in press as: A. Kaya, A.B. Can, A weighted rule based method for predicting malignancy of pulmonary nodules by nodule characteristics, J Biomed Inform (2015), http://dx.doi.org/10.1016/j.jbi.2015.05.011