COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE 136 (2016) 97–106

journal homepage: www.intl.elsevierhealth.com/journals/cmpb

Correlation coefficient based supervised locally linear embedding for pulmonary nodule recognition

Panpan Wu [a,b], *Kewen Xia* [a], *Hengyong Yu* [b,*]

[a] *School of Electronic and Information Engineering, Hebei University of Technology, Tianjin, 300401, China*
[b] *Department of Electrical and Computer Engineering, University of Massachusetts Lowell, Lowell, MA 01854, USA*

ARTICLE INFO

Article history:
Received 19 November 2015
Received in revised form
1 July 2016
Accepted 11 August 2016

Keywords:
Dimensionality reduction
Supervised locally linear embedding
Spearman's rank correlation coefficient
Pulmonary nodule recognition

ABSTRACT

Background and objective: Dimensionality reduction techniques are developed to suppress the negative effects of high dimensional feature space of lung CT images on classification performance in computer aided detection (CAD) systems for pulmonary nodule detection.
Methods: An improved supervised locally linear embedding (SLLE) algorithm is proposed based on the concept of correlation coefficient. The Spearman's rank correlation coefficient is introduced to adjust the distance metric in the SLLE algorithm to ensure that more suitable neighborhood points could be identified, and thus to enhance the discriminating power of embedded data. The proposed Spearman's rank correlation coefficient based SLLE (SC^{2}SLLE) is implemented and validated in our pilot CAD system using a clinical dataset collected from the publicly available lung image database consortium and image database resource initiative (LICD-IDRI). Particularly, a representative CAD system for solitary pulmonary nodule detection is designed and implemented. After a sequential medical image processing steps, 64 nodules and 140 non-nodules are extracted, and 34 representative features are calculated. The SC^{2}SLLE, as well as SLLE and LLE algorithm, are applied to reduce the dimensionality. Several quantitative measurements are also used to evaluate and compare the performances.
Results: Using a 5-fold cross-validation methodology, the proposed algorithm achieves 87.65% accuracy, 79.23% sensitivity, 91.43% specificity, and 8.57% false positive rate, on average. Experimental results indicate that the proposed algorithm outperforms the original locally linear embedding and SLLE coupled with the support vector machine (SVM) classifier.
Conclusions: Based on the preliminary results from a limited number of nodules in our dataset, this study demonstrates the great potential to improve the performance of a CAD system for nodule detection using the proposed SC^{2}SLLE.

© 2016 Elsevier Ireland Ltd. All rights reserved.

1. Introduction

The aggressive and heterogeneous nature of lung cancer has made it a prominent concern in the war against cancer. Lung cancer is the second most common and the primary cause of cancer-related death in both men and women. In the United States, the estimated new cases and deaths in 2013 were 228,190 and 159,480, respectively [1]. It has been shown that computed tomography (CT) screening can improve early detection accuracy of lung cancer in high-risk individuals [2]. Therefore, early detection of potentially cancerous pulmonary nodules

* *Corresponding author.* Department of Electrical and Computer Engineering, University of Massachusetts Lowell, Lowell, MA 01854, USA. E-mail address: hengyong-yu@ieee.org (H. Yu).

http://dx.doi.org/10.1016/j.cmpb.2016.08.009
0169-2607/© 2016 Elsevier Ireland Ltd. All rights reserved.

becomes considerably crucial to improve the patients' relative survival rate. Significant efforts have been made to develop computer aided detection system for early detection of lung lesions from CT images [3–7]. A CAD system could significantly enhance the sensitivity and specificity of spiral CT lung screening and reduce costs by reducing physician time needed for interpretation. It is an alternative option for radiologists before suggesting a biopsy test [5].

The procedures of a CAD system mainly include CT image preprocessing, region of interest (ROI) extraction, feature extraction and classification. It is well known that feature extraction and classification are the two key steps and they have significant impacts on the effectiveness of the CAD system. Specifically, the input space of the pattern classifier will directly impact the classification performance. The complexity of medical characteristics in lung CT images determines high dimensional feature space to present pulmonary nodules, and plenty of redundancies and correlations hide important relationships between different feature variables. This might lead to negative effects on classification performance. Thus, dimensionality reduction (DR) techniques have been developed to eliminate the redundancies of the data to obtain more informative, descriptive, and compact data representations for subsequent classifications. This can also help to reduce the requirements of computational cost and memory and potentially enhance the discriminating power.

Dimensionality reduction methods can fall into two categories: feature selection and feature extraction. While both feature selection and extraction approaches result in some loss of information compared to the original raw data, they are effective ways to deal with high dimensional data for classification problems. Feature selection usually chooses feature subset directly from the original feature space based on certain criteria, while feature extraction obtains subset by projecting the original data to lower-dimensional intrinsic spaces. They also have received significant attention in lung nodule detection. Aoyama et al. [8] employed feature selection to choose different combinations of features by evaluating the performance of linear discriminant analysis (LDA) in distinguishing begin nodules from malignancy ones in terms of receiver operating characteristic (ROC) analysis. They added or removed features one-by-one in an iterative way and finally received the AUC (area under the ROC curve) value of 0.84 when multiple slices were used. In references [3,9], feature selection stage was carried out to determine the subset of candidate features based on the area under the ROC curve by using different classifiers. Nevertheless, these feature selection methods have drawbacks that cannot be negligible. They need complex computation to evaluate all the features, and it is difficult to avoid local optimum. Besides, they are often not robust in complex scenes. Thus, researchers attempt to use feature extraction approaches, which are more robust to variation. And they are computationally superior to the optimal feature selection methods [10]. Theoretically speaking, feature extraction is used to obtain meaningful low-dimensional structures latent in high-dimensional data. Classical approaches, such as principal components analysis (PCA) or multidimensional scaling (MDS), work well in linear cases. However, the intrinsic structures of real-world data are often highly nonlinear and cannot be approximated by linear manifolds. Recently, a promising solution is to use nonlinear manifold learning algorithms [11], i.e., locally linear embedding (LLE), Isomap, Laplacian Eigenmaps (LE). Those methods have a small number of free parameters, and they cannot be trapped by local minima, and the non-iterative form makes them simple to implement to obtain the embedding [12–15]. These methods are supposed to overcome the difficulties encountered by other classical nonlinear approaches (e.g., the self-organizing map, generative topographic mapping, mixtures of linear models, etc.). However, they are unsupervised and mostly intended for data mining and visualization when the number of classes and relationships between elements of different classes are unknown, and users often want to observe the data structure in order to make a decision about what to do next. As aforementioned, the goal of our CAD platform is to distinguish the true nodules from non-nodules, which is a two-category problem. The feature dataset contains two (often disjoint) manifolds, corresponding to two classes. To solve this problem, de Ridder et al. extended the concept of LLE to multiple manifolds and developed a supervised LLE (SLLE) algorithm which has been proven to be a suitable feature extraction step prior to classification [16].

The dissimilarity between data samples from different categories can be measured by their distances. It is generally believed that the neighborhood of a sample from one class should be composed of samples belonging to the same class. In the SLLE, by taking into account label information, the inter-class distance is greater than the Euclidean distance by adding a constant to the pairs of points belonging to different classes. Otherwise, it remains as the Euclidean distance. It has been demonstrated that the SLLE is a powerful feature extraction method, which can yield promising recognition results coupling with simple classifiers. Subsequently, various improved SLLE methods were proposed to enhance the performance of SLLE. Liu et al. [17] proposed a new SLLE in tensor space (SLLE/T) where a local manifold structure within the same class is preserved and the separability between different classes is enforced by maximizing distance of each point with its neighbors. Wen and Jiang [18] designed a rescaling distance function to shrink the intra-class distance and kept the inter-class distance. Zhang [19] modified the distance metric by shrinking the intra-class distance while expanding the inter-class distance to strengthen the discriminating power and generalization ability of embedded results in dimensionality reduction. Experimental results demonstrated that the improved distance method can yield better classification performance on lung nodule classification [10]. Similarly, a kernel Euclidean distance was introduced by Zhou et al. [20] to define the distance metric to map the data into feature space where points belonging to the same classes are close to each other while points belonging to different classes are far away from each other. Zhao and Zhang [21] designed a probability-based distance metric that enlarges the Euclidean distance for labeled and unlabeled points. The enlarged quantity of the distance is variable and proportional to the probability of two points belonging to different classes. However, the aforementioned literatures did not take it into account that the Euclidean distance merely considers the intrinsic geometry of the data [22]. The Euclidean distance could not well represent the similarity between data points in high dimensional space, which might lead to undesirable neighborhood.

Download English Version:

https://daneshyari.com/en/article/4958283

Download Persian Version:

https://daneshyari.com/article/4958283

Daneshyari.com