Published in final edited form as:

Comput Methods Programs Biomed. 2016 November ; 136: 97–106. doi:10.1016/j.cmpb.2016.08.009.

Correlation Coefficient based Supervised Locally Linear Embedding for Pulmonary Nodule Recognition

Panpan Wu[1,2], **Kewen Xia**[1], and **Hengyong Yu**[2,*]

[1]School of Electronic and Information Engineering, Hebei University of Technology, Tianjin, 300401, China

[2]Department of Electrical and Computer Engineering, University of Massachusetts Lowell, Lowell, MA 01854, United States

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^2SLLE) 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^2SLLE, as well as SLLE and LLE algorithm are applied to reduce the dimensionality. Several quantitative measurements are also used to evaluate and compare the performance.

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^2SLLE.

*Corresponding Author: hengyong-yu@ieee.org.

Keywords

Dimensionality reduction; supervised locally linear embedding; Spearman's rank correlation coefficient; pulmonary nodule recognition

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 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 [8].

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. Masahito A. et al [9] 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 [10] and [11], 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 [12]. Theoretically speaking, feature extraction is 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 [13], i.e., locally linear embedding (LLE), Isomap, Laplacian Eigenmaps (LE). Those methods have a small number of free parameters, they cannot be trapped by local minima, and the non-iterative form makes them simple to implement to obtain the embedding [14–17]. 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 proved to be a suitable feature extraction step prior to classification [18].

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 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.* [19] 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 [20] designed a rescaling distance function to shrink the intra-class distance and kept the inter-class distance. Zhang [21] 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 [12]. Similarly, a kernel Euclidean distance was introduced by

Zhou [22] 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 [23] 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 [24]. The Euclidean distance could not well represent the similarity between data points in high dimensional space, which might leads to undesirable neighborhood.

Inspired by the aforementioned work, in this paper, we propose to perform dimensionality reduction for the input space of the pattern classifier before the classification procedure using an improved Spearman's rank correlation coefficient (hereinafter referred to as Spearman correlation coefficient) based SLLE algorithm (SC^{2}SLLE). Unlike the classic SLLE that mainly focus on maintaining the intra-class distances and enlarging the inter-class distances, the improved algorithm not only enlarges the inter-class dissimilarity but also adjusts both the intra- and inter-class dissimilarity by modifying the distance metric with the correlation coefficients. Because the improved algorithm can help to find more suitable neighborhood, the discriminating power of embedded data for multi-class issues is enhanced and it can help to detect cancerous nodules with a high probability at the early stage. The proposed algorithm is also evaluated on a clinical data set downloaded from the publicly available lung image database consortium and image database resource initiative (LIDC-IDRI) database. Experiment results demonstrate the promising performances.

The rest of this paper is organized as follows. In section 2, the LLE and SLLE algorithm are briefly reviewed and the proposed SC^{2}SLLE is elaborated. In section 3, the detail experimental materials, procedure and results are presented. The related issues are discussed in section 4 and the conclusion is made in the last section.

2. Research Methodology

2.1. LLE

Suppose the input dataset $\mathbf{X}$ consists of N real-valued vectors $\mathbf{X} = [\mathbf{x}_1, \cdots \mathbf{x}_i, \cdots \mathbf{x}_N]$ where $\mathbf{x}_i \in \mathbf{R}^D$ and D is the dimension of the vectors. The output $\mathbf{Y}$ of LLE, $\mathbf{Y} = [\mathbf{y}_1, \cdots \mathbf{y}_i, \cdots \mathbf{y}_N]$ with $\mathbf{y}_i \in \mathbf{R}^d$, is obtained by mapping the high dimensional input $\mathbf{X}$ into a single global coordinate system of lower dimensionality d ($d \ll D$). The classic LLE [16, 17] mainly includes the following three steps.

S1 Finding the K neighbors of each data point $\mathbf{x}_i$ in terms of the Euclidean distance. We denote the indexes of the neighborhood of each data $\mathbf{x}_i$ as $\mathbf{J}_i$, and the neighbors of $\mathbf{x}_i$ are defined as $\mathbf{x}_j, j \in \mathbf{J}_i$.

S2 Computing the reconstruction weights $\mathbf{W}_i$ by minimizing the reconstruction errors

$$\varepsilon(\mathbf{W})=\sum_{i=1}^{N}\left\|\mathbf{x}_i-\sum_{j=1}^{N}w_{ij}\mathbf{x}_j\right\|^2 \quad (1)$$

subject to two constraints

$$s.t.\ \sum_{j=1}^{N}w_{ij}=1,\ \ w_{ij}=0\ \ j\notin\mathbf{J}_i. \quad (2)$$

The weights w_{ij} summarize the contribution of the j^{th} data point to the i^{th} reconstruction and the optimal weights can be found by solving a least-squares problem. By exploiting the constraints in Eq. (2), the reconstruction error can be rewritten as

$$\varepsilon(\mathbf{W}_i)=\left\|\mathbf{x}_i-\sum_{j=1}^{K}w_{ij}\mathbf{x}_j\right\|^2=\left\|\sum_{j=1}^{K}w_{ij}\mathbf{x}_i-\sum_{j=1}^{K}w_{ij}\mathbf{x}_j\right\|^2=\left\|\sum_{j=1}^{K}w_{ij}(\mathbf{x}_i-\mathbf{x}_j)\right\|^2. \quad (3)$$

Denoting the local covariance matrix $\mathbf{Q}_{jl}=(\mathbf{x}_i-\mathbf{x}_{ij})^T(\mathbf{x}_i-\mathbf{x}_{il})$, we have

$$\varepsilon(\mathbf{W}_i)=\sum_{j=1}^{K}\sum_{l=1}^{K}w_{ij}\,w_{il}\,\mathbf{Q}_{jl}. \quad (4)$$

The reconstruction error Eq. (4) can be minimized in a closed form using a Lagrange multiplier in terms of the inverse local covariance matrix

$$F(\mathbf{W}_i)=\sum_{j=1}^{K}\sum_{l=1}^{K}w_{ij}w_{il}\mathbf{Q}_{jl}+\lambda\left(\sum_{j=1}^{K}w_{ij}-1\right), \quad (5)$$

and the optimal weights can be calculated as

$$w_{ij}=\frac{\sum_{l=1}^{K}\mathbf{Q}_{jl}^{-1}}{\sum_{p=1}^{K}\sum_{q=1}^{K}\mathbf{Q}_{pq}^{-1}}. \quad (6)$$

it should be noticed that when the neighbor number is greater than the input dimension ($K>D$), the local covariance matrix is singular or nearly singular, and it can be solved by adding a regularization term

$$\mathbf{Q}_{jl}=\mathbf{Q}_{jl}+\gamma Tr\left(\mathbf{Q}_{jl}\right)\mathbf{I}_{jl}, \quad (7)$$

where γ is a regularization parameter whose value is usually set between $[10^{-3},10^{-5}]$, $Tr(\mathbf{Q}_{jl})$ is the trace of $\mathbf{Q}_{jl}$ and $\mathbf{I}_{jl}$ is the identity matrix.

S3 Computing the low-dimensionality embedding $\mathbf{y}_i$ that best preserve the local intrinsic geometric properties hidden in the high-dimensional space by fixing the weights $\mathbf{W}_i$ and optimizing the coordinates $\mathbf{y}_i$. The embedding vectors $\mathbf{y}_i$ can be found by minimizing the cost function:

$$\phi(\boldsymbol{Y})=\sum_{i=1}^{N}\left\|\mathbf{y}_i-\sum_{j=1}^{N}w_{ij}\mathbf{y}_j\right\|^2, \quad (8)$$

$$s.t.\ \sum_{i=1}^{N}\mathbf{y}_i=0,\ \frac{1}{N}\sum_{i=1}^{N}\mathbf{y}_i\mathbf{y}_i^T=\mathbf{I}, \quad (9)$$

where $\mathbf{I}$ represents the unit identity matrix. The two constrains make the coordinates of $\mathbf{y}_i$ centered on the origin with a unit covariance. The cost function is further transformed into the following form,

$$\phi(\boldsymbol{Y})=\sum_{\mathrm{i}=1}^{N}\left\|\boldsymbol{Y}\mathbf{I}_i-\boldsymbol{Y}\mathbf{w}_i\right\|^2=\sum_{\mathrm{i}=1}^{N}\left\|\boldsymbol{Y}(\mathbf{I}_i-\mathbf{w}_i)\right\|^2. \quad (10)$$

Based on the properties of matrix transformation, the cost is written as,

$$\phi(\boldsymbol{Y})=tr\left(\mathbf{Y}\mathbf{M}\boldsymbol{Y}^T\right) \quad (11)$$

where $\mathbf{M}=(\mathbf{I}-\mathbf{W})^T(\mathbf{I}-\mathbf{W})$ is a $N\times N$ matrix. With the constraints in Eq. (9), this optimization problem Eq.(11) can be solved by the Lagrange multiplier. Then we have,

$$\mathrm{M}\boldsymbol{Y}^T=\lambda\boldsymbol{Y}^T. \quad (12)$$

Thus the optimal embedding can be found by computing the bottom $d+1$ eigenvectors of the sparse matrix $\mathbf{M}$. The bottom eigenvector of this matrix is the unit vector with all equal components, representing a free translation mode of eigenvalue zero. It is discarded to enforce the constraint $\sum_{i=1}^{N}\mathbf{y}_i=\mathbf{0}$ in Eq. (9), and the remaining d eigenvectors yield the final embedding $\mathbf{Y}$.

2.2. SLLE

The supervised LLE was introduced to deal with data sets containing multiple (often disjoint) manifolds corresponding to different classes [18]. For each data point $\mathbf{x}_i$, it should be reconstructed from its neighbors belonging to the same class for the purpose of

classification. One way is to enlarge the Euclidean distance by adding a constant to the pairs of points from different classes, and the distance of data points from the same class is kept. Mathematically, it can be expressed as

$$O(i,j)=o(i,j)+\alpha \max\left(\{o(i,j)\}\right)\Delta_{ij}, \quad (13)$$

where $o(i, j)$ is the Euclidean distance between $\mathbf{x}_i$ and $\mathbf{x}_j$, max ($\{o(i, j)\}$) is the maximum value of the Euclidean distance set $\{o(i, j)\}$ between data points, $\Delta_{ij} = 1$ when $\mathbf{x}_i$ and $\mathbf{x}_j$ belong to different classes, $\Delta_{ij} = 0$ when $\mathbf{x}_i$ and $\mathbf{x}_j$ belong to the same class, and $\alpha \in [0,1]$ is a tuning parameter. When $\alpha = 0$, the SLLE is degraded to the original unsupervised LLE; When $\alpha = 1$, one obtains the fully supervised LLE (1-LLE); A varying α between 0 and 1 gives a partially supervised LLE (α-LLE).

2.3. Spearman Correlation Coefficient based SLLE (SC^{2}SLLE)

An ideal neighborhood searching mechanism should attempts to maximize the inter-class dissimilarity and minimize the intra-class dissimilarity. Based on this principle, the SLLE algorithm is employed rather than the LLE. In this paper, the Spearman correlation coefficient is introduced to further improve the performance of SLLE, which is named SC^{2}SLLE. The SLLE is typically implemented by using the Euclidean distance. However, the Euclidean distance may assign a data point neighbors that in fact are far away [24]. The SC^{2}SLLE algorithm is developed by modifying the neighborhood searching mechanism in the SLLE. In statistics Spearman correlation coefficient is a nonparametric measure of rank correlation. It assesses how well the relationship between two variables (whether linear or not) can be described using a monotonic function [25], and it is not necessary to assume the distribution of the data. It is also independent of the spatial geometry of the data, thus it can be employed to search the neighbors of strong correlation with a given data point.

The significant difference between the SC^{2}SLLE and SLLE lies in the neighborhood searching mechanism. The SLLE enlarges the Euclidean distance by adding a constant to the pairs of points from different classes while others are kept unchanged. Let ρ_{ij} be the Spearman correlation coefficient between $\mathbf{x}_i$ and $\mathbf{x}_j$. Because a greater $|\rho_{ij}|$ implies stronger correlation between $\mathbf{x}_i$ and $\mathbf{x}_j$ and a smaller Euclidean distance $o(i, j)$ means better similarity, we combine the two different measures to achieve stronger capability to search the nearest neighbors in high dimensional space. We use $1 - |\rho(i, j)|$ to multiply the Euclidean distance $o(i, j)$ to achieve the goal. The modified distance metric is defined as:

$$O^{cc}(i,j)=(1-|\rho(i,j)|)o(i,j)+\alpha \max\left(\{1-|\rho(i,j)|\}\, o(i,j)\right)\Delta_{ij}, \quad (14)$$

Based on the new distance formula Eq. (14), we have the SC^{2}SLLE algorithm as follows. First, we find the K nearest neighbors of each data point $\mathbf{x}_i$ in terms of the distance metric Eq. (14); then, we follow the same procedures S2 and S3 for LLE in section 2.1.

2.4 SC^{2}SLLE for Classification

When the supervised dimension reduction algorithm is applied for classification, the training and testing samples need to be considered separately. After the dimension reduction, the training data with label information can be used to train a classifier. However, there is no label information in the testing set. While the steps of SC^{2}SLLE algorithm in 2.3 can be performed to obtain the embedding results for the training set, they should be modified for the testing set as follows.

S1 For each data $\mathbf{x}_i$ in the testing set, finding the K nearest neighbors $\mathbf{x}_j (j = 1, \cdots K)$ in the training set in terms of the adjusted Euclidean distance

$$o^{cc}(i,j) = (1 - |\rho(i,j)|) o(i,j); \quad (15)$$

S2 Computing the weight $\mathbf{W}_i$ that best reconstruct each testing data $\mathbf{x}_i$ from its neighbors;

S3 Interpolating $\mathbf{y}_i$ from the corresponding $\mathbf{y}_j (j = 1, \cdots K)$ in the training set using the weight $\mathbf{W}_i$

$$\mathbf{y}_i = \sum_{j=1}^{K} w_{ij} \mathbf{y}_j. \quad (16)$$

3. Experiments and Results

3.1. Data Preparation

In this pilot study, the goal is to distinguish the pulmonary nodules from the non-nodules, which is a two-category classification problem. A set of clinical images was downloaded from the lung image database consortium (LICD) and image database resource initiative (IDRI). The specific objective of the LIDC was to provide a reference database for the relative evaluation of image processing or CAD algorithms [26]. The LIDC-IDRI database (http://ncia.nci.nih.gov/ncia/) consists of 1018 cases. Each case includes images from a clinical thoracic CT scan and an associated extensible markup language (XML) file that records the results of a two-phase annotation process performed by four experienced thoracic radiologists. It is a web-accessible international resource for development, training, and evaluation of CAD methods for lung cancer detection and diagnosis [27].

Because solidary pulmonary nodules are the dominating type of nodule in the whole database, we focus on the solitary pulmonary nodules in this pilot study. Two rules are considered to select the training and testing sets. On the one hand, the associated XML files are employed to ensure that correct nodule type is chosen. For nodules > =3mm, each reader is asked to subjectively assess the nodule's several characteristics in a 1–5 scale, such as subtlety, internal structure, speculation, texture, malignancy, etc. Two of them provide information about solid nodules: one is internal structure, which means the internal structure or expected internal composition of the nodules (1: soft tissue, 2: fluid, 3: fat, 4: air); and the

other is texture, which is defined by 3 terms (1:non-solid/ground glass opacity, 3: part solid/ mixed, 5: solid texture) [28]. Those characteristics are recorded in the associated XML files. In term of these two characteristics, we can narrow down the search space and select the images of primary interest. On the other hand, because the LIDC annotations provide no information on the nodule typology, it is based on the visual assessment to determine whether the nodules used in this paper are juxta-pleural.

In our experiments, 60 cases were randomly collected based on the aforementioned rules to evaluate the effectiveness of the proposed algorithm. After a series of medical image processing steps presented in our previous work [29], 204 candidate nodules of size > 3mm were extracted, consisting of 64 nodules and 140 non-nodules marked by at least one radiologist.

3.2. Experimental Design

The flowchart of our CAD system for pulmonary nodules detection is shown in Fig. 2. Our experiment includes five major steps.

S1 **Image Preprocessing**. Image de-noising and pulmonary parenchyma segmentation are carried out. Median filter is performed to suppress the Gaussian noise in the CT images. The well-known thresholding method is employed to perform an initial segmentation in terms of the grayscale values between pulmonary parenchyma and surrounding areas. After that, other image analysis techniques (e.g. filling, region growing, morphology operation, multiply operator) are implemented to remove the background and the interference tissues (e.g. bronchus and blood vessels) to obtain the complete pulmonary parenchyma. Longitudinal scan and morphological erosion operation are also explored when there are connections between the left and right pulmonary parenchyma.

S2 **Region of Interest (ROI) Extraction**. All the suspicious nodules are extracted in this step. An optimal thresholding algorithm is executed to determine the preliminary ROIs according to the difference of CT values among the lung parenchyma and lesions. Classical image processing technology and circular filter are designed to eliminate the highlights noise points and suppress the linear structures, respectively.

S3 **Feature Extraction**. Thirty-four features are extracted from the ROIs, which are obtained in the second step depending on the manifestation of solidary pulmonary nodules in CT images. More details about extracted nodule features can be found in our previous work [28]. All those extracted features are listed in Table 1. They are gray features including gray mean and gray variance, morphological features such as seven invariant moments, area, diameter, long and short axis, circularity and compactness and texture features consisting of contrast, correlation, angular second moment and homogeneity based on gray-level co-occurrence matrix (GLCM) along four directions 0°, 45°, 90° and135°, thus 4 × 4= 16 texture features are included.

S4 **Feature Dimensionality Reduction**. It is the key step that this paper focuses on. The features extracted from step 3 are high-dimensional data, and they may include plenty of redundancies and correlations hiding important relationships. Therefore, our proposed SC^{2}SLLE algorithm is utilized to reduce the dimensionality, aiming to eliminate the redundancies and obtain more meaningful low-dimensional structures hidden in high dimensional data. This is helpful for subsequent classification operations.

S5 **Classifier Design.** Finally, to further eliminate false positive nodules, support vector machine (SVM) based classifiers are trained and used for classification.

Note: In table 1, we assume an $I \times J$ image with the pixel intensity $f(i, j)$ corresponding to the coordinate (i, j). $\eta_{pq}=\mu_{pq}/\mu_{00}^{r}$, $r = 1 + (p + q)/2$ is the normalized central moment with order (p, q), where $\mu_{pq} = \Sigma_i \Sigma_j (i - \bar{\imath})^p (j - \bar{\jmath})^q f(i, j)$ is the (p, q) order central moment, $\bar{\imath} = m_{10}/m_{00}$, $\bar{\jmath} = m_{01}/m_{00}$ and $m_{pq} = 2\Sigma_i \Sigma_j i^p j^q f(i, j)$. $p(i, j)$ is the probability density of the gray-level co-occurrence matrix (GLCM).

3.3. Results and Analysis

To evaluate the embedding performance of the proposed method, not only the SC^{2}SLLE but also the LLE and SLLE are implemented for dimension reduction. Based on the observation that all the raw features have different scales, a normalization step is employed to make all the features in a common scale. This can help to eliminate the effects caused by different scales. In this study, the raw features are standardized by using the standard scores as follow,

$$x^* = \frac{x - \mu}{\sigma}, \quad (17)$$

where x is a given raw feature, μ is the mean of the feature population, σ is the standard deviation of the feature population, and x^* is the corresponding normalized feature.

The first step for dimensionality reduction techniques is parameter setting. In the LLE, there are two parameters, which are embedded dimension d and the number of neighbors for each data point k. In the SLLE and SC^{2}SLLE, there are three parameters: the embedded dimension d, the number of neighbors k and the tuning parameter α. Mapping quality is quite sensitive to these parameters. For the embedded dimension d, on one hand, if d is too big, the mapping will enhance noise; and on the other hand, if d is too small, distinct parts of the data set might be mapped on the top of each other. Therefore, various automatic techniques have been developed to estimate the intrinsic dimensionality of a given dataset. In this experiment, we employed several state-of-the-art intrinsic dimensionality estimation techniques, such as MLE, MiND_ML, MiND_KL, DANCo and DANCoFit [30, 31] to determine the value d. As can be seen in Fig. (2), the majority of these methods move towards agreement around $d = 10$, thus we employ $d = 10$ for our system. Regarding the parameter k, the mapping will not reflect the global properties if k is too small; if it is too big, the mapping will lose its nonlinear character because the entire data set is seen as a local neighborhood. Given that there are 51(or 52) nodules and 112 non-nodules in the training

set, when $k \leq 52$, it can be guaranteed that those k neighbors could be chosen from the same class. Otherwise, the k neighbors come from different classes, and one could not make full use of the label information of the training set which will diminish the advantages of the SLLE and SC^{2}SLLE. Hence, this process is implemented empirically by investigating the range of $12 \leq k \leq 52$ with an interval 1. The parameter α controls the amount to which the class information should be incorporated [32]. Here, $0 \leq \alpha \leq 1$ with an interval 0.1 is performed.

To quantitatively evaluate the effectiveness of the proposed SC^{2}SLLE, the SVM classifier is trained and employed, with the purpose of reducing the false positive nodules and identifying the true ones. The penalty parameter c of the error term and the Gauss kernel parameter γ in the SVM are automatically determined through a grid search approach for each experiment [33]. A standard tool in statistics known as N-fold cross-validation (CV) is adopted to improve the credibility of classification results. We take N= 5, so five experiments will be executed and the results are averaged. The feature dataset is randomly divided into five different subsets equally. In one experiment, four subsets are used to train the classifier and the remaining one is assigned as the test dataset for validation. This experiment is repeated in 5 different ways, and each subset is unbiasedly evaluated once. Table 2 lists all the possible outcomes of a test procedure and the gold standard.

A large number of experiments are conducted to validate the performance of the algorithms. Table 3 summarizes the optimal values α and k for the best classification accuracy with respect to different dimensionality reduction technique. As shown in Table 3, the best classification accuracy (92.68%) is achieved by the SC^{2}SLLE when α= 0.2 and k= 23, while the counterparts of the SLLE and LLE are 87.80 % (α= 0.2, k= 28,30) and 80.49 %(k= 19). Notice that only the first occurrence of best result is recorded in Table 3 in each experiment. On average, the mean accuracies for the three methods are 87.65%, 84.30% and 76.35%, respectively. We can see that the supervised SC^{2}SLLE and SLLE outperform the unsupervised LLE in terms of the classification accuracy. Moreover, to achieve the best classification accuracy, the parameters α and k should be set neither too small nor too big due to their sensitivity.

To comprehensively evaluate the performance of CAD systems for lung nodule detection, in addition to the classification accuracy, several metrics stand out based on the outcomes of the validation procedure: sensitivity, specificity, false positive rate (FPR) and ROC analysis. Performance comparisons are summarized in Table 4. Experimental results of nodule recognition without employing dimensionality reduction method are also presented in Table 4 for comparison. As can be seen in Table 4, the LLE algorithm doesn't show any advantage compared to the situation when the dimensionality reduction process is not applied in the system, and it is even worse than that in terms of accuracy and sensitivity. However, clear performance improvements can be observed when the SLLE and SC^{2}SLLE algorithm are utilized. One can also compare the performances of SC^{2}SLLE and SLLE algorithms. It is shown that better classification accuracy can be obtained by the SC^{2}SLLE than the SLLE. Both of them have a comparable specificity and false positive rate. Particularly, better sensitivity (79.23%) is achieved by the SC^{2}SLLE compared to the SLLE (68.97%), which is of crucial importance in clinical applications.

Furthermore, p-values between different performances of the LLE, SLLE and SC^{2}SLLE are computed using t-test. Fig. 4 shows the overall performance with different approaches in our CAD system in terms of accuracy, AUC, and p-value. It can be observed that the differences between performances regarding LLE vs. SLLE, as well as LLE vs. SC^{2}SLLE are significant since the p-values are all less than 0.05 (significant level $p = 0.05$) in terms of both accuracy and AUC. Although no significant performance differences ($p>0.05$) are observed between the SLLE and SC^{2}SLLE, it is evident that SC^{2}SLLE is superior to SLLE ($p<0.25$) based on the average classification accuracy and the average AUC. Our analysis indicates that the supervised dimensionality reduction algorithm boosts the classification performance and the proposed SC^{2}SLLE outperforms the SLLE algorithm.

4. Discussion

In this paper, an improved SLLE algorithm is proposed by incorporating the correlation coefficient, named SC^{2}SLLE. By combining the two different measures to search the optimal nearest neighbors, the SC^{2}SLLE modifies the dissimilarity between data points by adjusting their Euclidean distance, in addition to enlarge the inter-class dissimilarity. The proposed algorithm can help to find more suitable neighborhood even for the data points in the same classes. Consequently, the discriminating power of embedded data is enhanced for multi-class issues. The proposed SC^{2}SLLE algorithm is employed in feature extraction in the framework of our pilot CAD system for lung nodule detection, and its effectiveness is demonstrated on nodules features extracted from the LIDC-IDRI database.

Because it is still a challenging problem to extract features for classification in CAD systems for lung nodule detection, extensive interests have been attracted to this area. However, it is difficult to make an objective comparison with previously published CAD systems due to the variability in the dataset (such as number of cases, scanning protocols), nodule type and size criterion, and different validation procedures. Nevertheless, we believe it is still important to attempt a relative comparison. For this purpose, we identified several representative methods that have reported better results. Li et al. [12] used supervised manifold learning algorithm to extract features before the lung nodule classification based on fusion of all-class and pairwise-class structure. Comparable classification results are obtained when it is combined with only one classifier, while the recognition accuracy was improved significantly when multi-classifiers system was employed. Nevertheless, the limitations lie in that different dataset was adopted and only classification accuracy was reported in their experiment. It is noteworthy that ensembles of alternative classifiers are possible solutions to maximize performance. As the aforementioned, in addition to the feature extraction, feature selection is also widely used in CAD systems in nodule detection. Zhu et al. [34] adopted a genetic algorithm to find and select textures features of solitary pulmonary nodules, and its performances were evaluated based on the selected features. A similar AUC value 0.8748 was presented by using the SVM classifier. However, the finally selected features are determined after 300 genetic generations, which requires ultra-high computational cost for clinical applications. Temesguen et al. [35] used a sequential forward selection process to determine the optimal feature subset. A 7-fold cross-validation performance analysis using the LIDC database showed a CAD sensitivity of 82.66% with an average of 3FPs per CT

scan/case. Nevertheless, 80.4% nodule candidates were correctly identified using 40 selected features in their CAD system, which is lower than that in our study.

It should be pointed out that the proposed algorithm also has some limitations. A lot of distance metrics have been proposed to enhance the performance of SLLE. Along this direction, more reasonable and potentially powerful distance formulas could be designed for multi-manifold space based on three basic principles: non-negativity, symmetric and triangle inequality. Besides, in this pilot study, we merely take the solitary pulmonary nodules into account without considering more sophisticated cases, and the entire feature extraction process is based on the traditional image processing techniques for the CT images. When we deal with different type of nodules or different database, the proposed algorithm is lack of generalization. From the reported results, though the proposed algorithm outperforms the original LLE and SLLE methods, no dramatic performance improvement is observed by our entire CAD system. This probably is due to lack of advanced image processing techniques or only with a small number of dataset. Thus, further research is needed to improve the existing systems for better solutions. The recently prevalent dictionary learning framework makes it possible to learn class-specific dictionaries and features directly from the original CT images. This can help to avoid the image processing steps to further improve the accurate rates. Meanwhile, we will try to extract more discriminating and compact features for specific classes with these class-specific dictionaries. Currently, the supervised classifier SVM is trained and a testing sampling is interpolated in the low dimensional space for classification. However, plenty of other information is available for many clinical applications, such as the class sizes of the data sets, the desired sensitivity for each class, *etc*. In the near future, we will incorporate this information into the training and classification procedures to further improve the performance of the CAD system.

5. Conclusion

In summary, an improved SLLE algorithm (SC^2SLLE) is proposed by incorporating the correlation coefficient for multi-classification problem. The proposed algorithm is employed in our pilot CAD system for lung nodule detection and a detailed performance comparison and analysis are presented based on the publicly available LIDC-IDRI database. Better experimental results are obtained with the improved algorithm compared to that with the LLE and SLLE algorithms. This study demonstrates the potential for improving the performance of the CAD system in nodule detection with a high probability of being cancers at its early stage.

Acknowledgments

LIDC-IDRI Attribution:

This work was supported in part by National Institute of Biomedical Imaging and Bioengineering R21 grant EB019074. The authors acknowledge the National Cancer Institute and the Foundation for the National Institutes of Health, and their critical role in the creation of the free publicly available LIDC/IDRI Database used in this study.

References

1. Siegel, Rebecca; Naishadham, Deepa; Jemal, Ahmedin. Cancer statistics, 2013. CA: a cancer journal for clinicians. 2013; 63(1):11–30. [PubMed: 23335087]
2. National Lung Screening Trial Research Team. Reduced lung-cancer mortality with low-dose computed tomographic screening. The New England journal of medicine. 2011; 365(5):395. [PubMed: 21714641]
3. Messay, Temesguen; Hardie, Russell C.; Rogers, Steven K. A new computationally efficient CAD system for pulmonary nodule detection in CT imagery. Medical Image Analysis. 2010; 14(3):390–406. [PubMed: 20346728]
4. Firmino, Macedo, et al. Computer-aided detection system for lung cancer in computed tomography scans: Review and future prospects. Biomed Eng Online. 2014; 13:1–16. [PubMed: 24410918]
5. Kuruvilla, Jinsa; Gunavathi, K. Lung cancer classification using neural networks for CT images. Computer methods and programs in biomedicine. 2014; 113(1):202–209. [PubMed: 24199657]
6. Agarwal, Ritika; Shankhadhar, Ankit; Sagar, Raj Kumar. Detection of Lung Cancer Using Content Based Medical Image Retrieval. Advanced Computing & Communication Technologies (ACCT), 2015 Fifth International Conference on; IEEE; 2015.
7. da Silva Sousa; Ferreira, João Rodrigo, et al. Methodology for automatic detection of lung nodules in computerized tomography images. Computer methods and programs in biomedicine. 2010; 98(1): 1–14. [PubMed: 19709774]
8. Kuruvilla, Jinsa; Gunavathi, K. Lung cancer classification using neural networks for CT images. Computer methods and programs in biomedicine. 2014; 113(1):202–209. [PubMed: 24199657]
9. Aoyama, Masahito, et al. Computerized scheme for determination of the likelihood measure of malignancy for pulmonary nodules on low-dose CT images. Medical Physics. 2003; 30(3):387–394. [PubMed: 12674239]
10. Taşcı, Erdal; Uğur, Aybars. Shape and texture based novel features for automated juxtapleural nodule detection in lung cts. Journal of medical systems. 2015; 39(5):1–13. [PubMed: 25600193]
11. Messay, Temesguen; Hardie, Russell C.; Rogers, Steven K. A new computationally efficient CAD system for pulmonary nodule detection in CT imagery. Medical Image Analysis. 2010; 14(3):390–406. [PubMed: 20346728]
12. Li, Ying; Yu, Qian. Lung Nodule Classification Using Supervised Manifold Learning Based on All-Class. Computer Science & Service System (CSSS); 2012 International Conference on; IEEE; 2012.
13. Kawata, Yoshiki, et al. SPIE Medical Imaging. International Society for Optics and Photonics; 2015. Nonlinear dimensionality reduction of CT histogram based feature space for predicting recurrence-free survival in non-small-cell lung cancer.
14. Karbauskaitė, Rasa; Kurasova, Olga; Dzemyda, Gintautas. Selection of the number of neighbours of each data point for the locally linear embedding algorithm. Information technology and control. 2015; 36(4)
15. Saul, Lawrence K.; Roweis, Sam T. An introduction to locally linear embedding. 2000. unpublished. Available at: http://www.cs.toronto.edu/~roweis/lle/publications.html
16. Saul, Lawrence K.; Roweis, Sam T. Think globally, fit locally: unsupervised learning of low dimensional manifolds. The Journal of Machine Learning Research. 2003; 4:119–155.
17. Roweis, Sam T.; Saul, Lawrence K. Nonlinear dimensionality reduction by locally linear embedding. Science. 2000; 290(5500):2323–2326. [PubMed: 11125150]
18. de Ridder, Dick, et al. Artificial Neural Networks and Neural Information Processing ICANN/ ICONIP 2003. Springer; Berlin Heidelberg: 2003. Supervised locally linear embedding; p. 333-341.
19. Liu, Chang, et al. Supervised locally linear embedding in tensor space. Intelligent Information Technology Application, 2009. IITA 2009. Third International Symposium on; IEEE; 2009.
20. Wen, Guihua; Jiang, Lijun. Clustering-based locally linear embedding. Systems, Man and Cybernetics, 2006. SMC'06. IEEE International Conference on; IEEE; 2006.

21. Zhang, Shi-qing. Enhanced supervised locally linear embedding. Pattern Recognition Letters. 2009; 30(13):1208–1218.
22. Zhou, Yun, et al. Enhanced supervised Kernel Neighborhood Preserving Embedding for Radar HRRP Recognition. Journal of Convergence Information Technology. 2013; 8(3)
23. Zhao, Lingxiao; Zhang, Zhenyue. Supervised locally linear embedding with probability-based distance for classification. Computers & Mathematics with Applications. 2009; 57(6):919–926.
24. Varini, Claudio; Degenhard, Andreas; Nattkemper, Tim W. ISOLLE: LLE with geodesic distance. Neurocomputing. 2006; 69(13):1768–1771.
25. https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient.
26. http://imaging.cancer.gov/programsandresources/informationsystems/lidc
27. Armato, Samuel G., III, et al. The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. Medical physics. 2011; 38(2):915–931. [PubMed: 21452728]
28. https://wiki.cancerimagingarchive.net/display/Public/LIDC-IDRI
29. Wu, Pan-pan; Xia, Ke-wen; Heng-yong, Yu. Relevance Vector Machine Based Pulmonary Nodule Classification. Journal of Medical Imaging and Health Informatics. 2016; 6(1):163–169.
30. De Ridder, Dick; Duin, Robert PW. Tech Rep PH-2002-01. Pattern Recognition Group, Dept. of Imaging Science & Technology, Delft University of Technology; Delft, The Netherlands: 2002. Locally linear embedding for classification; p. 1-12.
31. Lombardi, Gabriele, et al. Minimum Neighbor Distance Estimators of Intrinsic Dimension. European Conference on Principles of Data Mining and Knowledge Discovery; 2011.
32. Ceruti, Claudio, et al. DANCo: An intrinsic dimensionality estimator exploiting angle and norm concentration. Pattern Recognition. 2014; 47(8):2569–2581.
33. Hsu, Chih-Wei; Chang, Chih-Chung; Lin, Chih-Jen. A Practical Guide to Support Vector Classification. Department of Computer Science, National Taiwan University; 2003.
34. Zhu, Yanjie, et al. Feature selection and performance evaluation of support vector machine (SVM)-based classifier for differentiating benign and malignant pulmonary nodules by computed tomography. Journal of digital imaging. 2010; 23(1):51–65. [PubMed: 19242759]
35. Messay, Temesguen; Hardie, Russell C.; Rogers, Steven K. A new computationally efficient CAD system for pulmonary nodule detection in CT imagery. Medical Image Analysis. 2010; 14(3)

Highlights

- A correlation coefficient is introduced to adjust the distance metric in the supervised locally linear embedding to ensure more suitable neighbors that could be chosen, and thus to enhance the discriminating power of embedded data.
- The method is validated on a clinical lung image database.

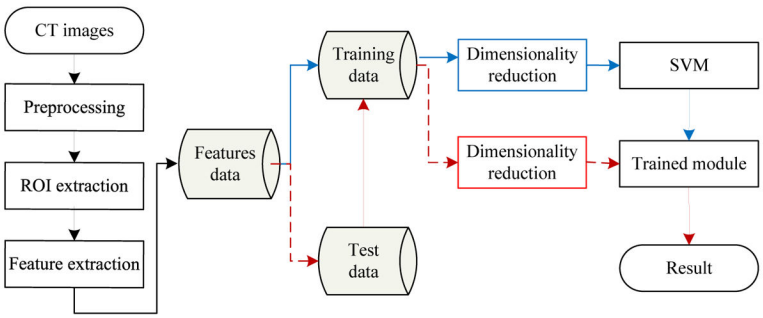

Figure 1.
The flowchart of our CAD system

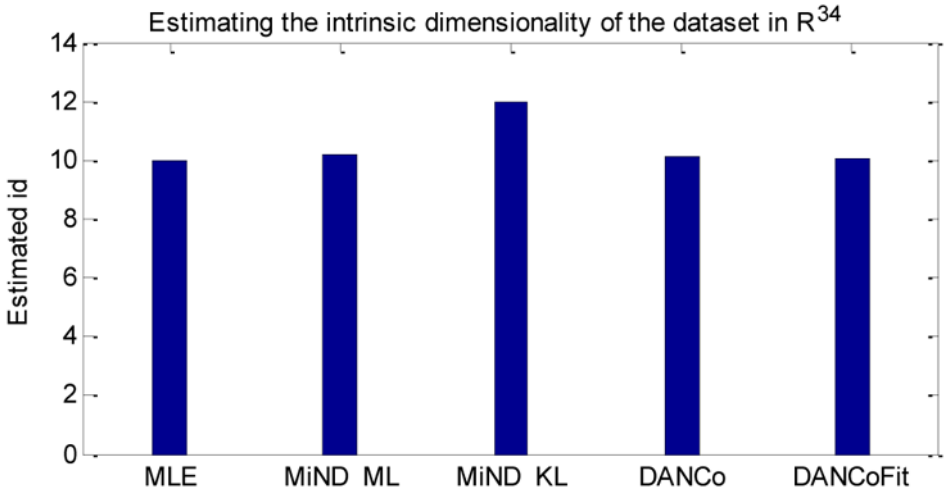

Figure 2.
Estimated intrinsic dimensionality with different techniques

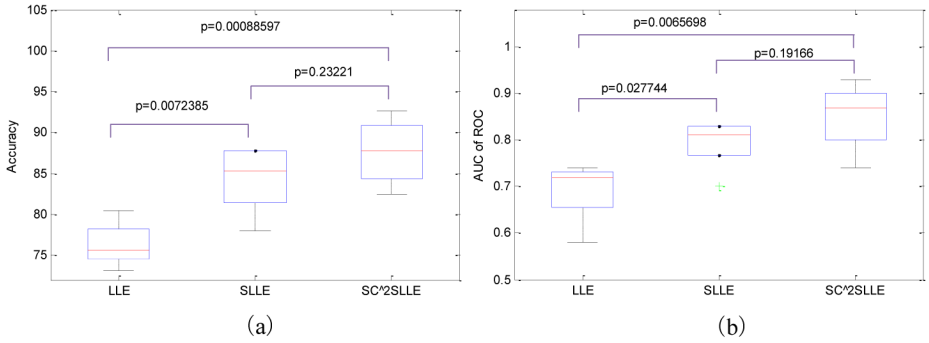

Figure 3.

Performance comparison of the classification system. (a) Accuracy v.s. p-value; and (b) AUC of ROC v.s. p-value.

Table 1

Features used in the experiment

name	feature
mean	$\frac{\sum_{i,j} f(i,j)}{I \times J}$
variance	$\frac{\sum_{i,j} (f(i,j) - mean)^2}{I \times J}$
seven invariant moments	$M_1 = \eta_{20} + \eta_{02}$
	$M_2 = (\eta_{20} + \eta_{02})^2 + 4\eta_{11}^2$
	$M_3 = (\eta_{30} + 3\eta_{12})^2 + (3\eta_{21} + \eta_{03})^2$
	$M_4 = (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2$
	$M_5 = (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2](3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - 2(\eta_{21} + \eta_{03})^2]$
	$M_6 = (\eta_{20} - \eta_{02})[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03})$
	$M_7 = (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2(\eta_{03} + 3\eta_{12})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]$
inscribed radius	R_i
circumradius	R_c
perimeter	P
long axis (l)	$2 \times \left[2(\mu_{20}+\mu_{02}+ \sqrt{(\mu_{20}-\mu_{02})^2+4\mu_{11}^2}/\mu_{00})\right]^{1/2}$
short axis	$2 \times \left[2(\mu_{20}+\mu_{02}- \sqrt{(\mu_{20}-\mu_{02})^2+4\mu_{11}^2}/\mu_{00})\right]^{1/2}$
circularity	R_i/R_c
compactness	$4\pi A/P^2$
constrast	$\Sigma_{i,j}(i-j)^2 p(i,j)$
correlation	$\sum_{i,j} \frac{(i-\mu_x)(j-\mu_y)p(i,j)}{\sigma_x \sigma_y}$
angular second moment	$\Sigma_{i,j} p^2(i,j)$
homogeneity	$\sum_{i,j} \frac{p(i,j)}{1+\lvert i-j \rvert}$
area (A)	The actual number of pixels in each ROI
flat	l/s

Table 2

All possible outcomes of a test

Test result	Gold standard	
	Positive	**Negative**
Positive	True Positive(TP)	False Positive(FP)
Negative	False Negative(FN)	True Negative(TN)
Total	TP+FN	FP+TN

Table 3

Optimal values of a and k for the best classification accuracy

5-fold CV	LLE		SLLE		SC^2 SLLE	
	k	Accuracy (%)	(a, k)	Accuracy (%)	(a, k)	Accuracy (%)
1	11	75.61	(0.2, 10)	78.05	(0.3, 11)	82.50
2	23	77.50	(0.2, 30)	87.80	(0.1, 25)	90.24
3	19	80.49	(0.2, 28)	87.80	(0.2, 23)	92.68
4	10	73.17	(0.4, 23)	85.37	(0.4, 14)	87.80
5	15	75.00	(0.3, 20)	82.50	(0.1, 21)	85.00

Table 4

Performance (mean) comparison with respect to different methods

Method	Accuracy (%) $\frac{TP+TN}{TP+TN+FP+FN}$	Sensitivity (%) $\frac{TP}{TP+FN}$	Specificity (%) $\frac{TN}{FP+TN}$	FPR (%) $\frac{FP}{FP+TN}$
Without DR	77.43	57.95	86.43	13.57
LLE	76.35	49.36	88.57	11.43
SLLE	84.30	68.97	91.43	8.57
SC^2SLLE	87.65	79.23	91.43	8.57