Journal of the Operations Research Society of China (2019) 7:515–537
https://doi.org/10.1007/s40305-019-00262-z

Layer-Wise Pre-Training Low-Rank NMF Model for Mammogram-Based Breast Tumor Classification

Wen-Ming Wu[1] · Xiao-Hui Yang[1] · Yun-Mei Chen[2] · Juan Zhang[3] · Dan Long[3] · Li-Jun Yang[1] · Chen-Xi Tian[1]

Received: 18 June 2018 / Revised: 3 November 2018 / Accepted: 2 August 2019 /
Published online: 1 October 2019

Abstract

Image-based breast tumor classification is an active and challenging problem. In this paper, a robust breast tumor classification framework is presented based on deep feature representation learning and exploiting available information in existing samples. Feature representation learning of mammograms is fulfilled by a modified nonnegative matrix factorization model called LPML-LRNMF, which is motivated by hierarchical learning and layer-wise pre-training (LP) strategy in deep learning. Low-rank (LR) constraint is integrated into the feature representation learning model by considering

This work was supported in part by the National Natural Science Foundation of China (No. 11701144), National Science Foundation of US (No. DMS1719932), Natural Science Foundation of Henan Province (No. 162300410061) and Project of Emerging Interdisciplinary (No. xxjc20170003).

✉ Xiao-Hui Yang
xhyanghenu@163.com

Wen-Ming Wu
wmwu55@163.com

Yun-Mei Chen
yun@ufl.edu

Juan Zhang
zhangjuan_496@163.com

Dan Long
legend_long@aliyun.com

Li-Jun Yang
yanglijun@henu.edu.cn

Chen-Xi Tian
tcxhaha@163.com

1 Data Analysis Technology Lab, Institute of Applied Mathematics, School of Mathematics and Statistics, Henan University, Kaifeng 475004, Henan, China

2 Department of Mathematics, University of Florida, Gainesville, FL 32611, USA

3 Zhejiang Cancer Hospital, Hangzhou 310022, China

the intrinsic characteristics of mammograms. Moreover, the proposed LPML-LRNMF model is optimized via alternating direction method of multipliers and the corresponding convergence is analyzed. For completing classification, an inverse projection sparse representation model is introduced to exploit information embedded in existing samples, especially in test ones. Experiments on the public dataset and actual clinical dataset show that the classification accuracy, specificity and sensitivity achieve the clinical acceptance level.

Keywords Breast tumor classification · Mammogram · LPML-LRNMF · Inverse space sparse representation · ADMM

Mathematics Subject Classification 68T10

1 Introduction

Breast tumor has become the most common malignant neoplasm for women. About 37.3% of breast tumor can be cured, especially in the case of early detection [1]. Effective breast tumor classification plays an important role in clinical diagnosis and treatment. The commonly used diagnostic techniques include mammography, magnetic resonance imaging (MRI) and near-infrared scanning [2]. Mammography is a common and effective breast tumor screening method [3], which can visualize nonpalpable and small tumors [4]. However, the performance of mammogram-based breast tumor classification may be decreased due to noise [5], and the distinction between cancerous and non-cancerous tumors may be subtle.

Feature extraction of mammograms will greatly improve the readability of these original data [4, 6–8]. Feature learning can further explore the more essential information. Deep learning is a popular feature representation learning method [9, 10]. Some preliminary results in recognizing benign and malignant tumor have been obtained [11]. However, the success of deep learning relies on complex network structures, high-performance GPU devices and optimized parallel algorithms. As a data-driven feature learning method, deep learning relies heavily on large number of effective training samples. However, tumor classification is a typical small sample problem. The nonnegative matrix factorization (NMF) is a feature learning method that does not pay attention to category information, and explores useful information contained in all available samples simultaneously, even if there are only a small number of training samples. In recent years, NMF [12] and its improved methods [13–19] have achieved good results for image-based tumor classification. Liu et al. [15] applied NMF to extract both appearance- and histogram-based semantic features of images. Li et al. [20] proposed a nonnegative low-rank matrix factorization (NLMF) method for image clustering. However, NMF is affected by the initial value of the iteration. Our previous work [21] proposed a layer-wise pre-training multilayer sparse NMF (LPML-SNMF) method by integrating NMF and deep representation learning. The LPML-SNMF is demonstrated effective for breast tumor classification based on microarray gene expression data, which has the characteristic of sparsity. It is certainly interesting and promising if we can complement advantages of different approaches.

For image-based tumor data, low rank (LR) is an important prior information for feature representation model. The optimization problem with low-rank regularization constraint is NP hard. A typical approach is to relax the problem by replacing the rank constraint with a nuclear norm l_* regularization [22]. There are many ways to optimize an l_* regularization problem [23–26]. The alternating direction method of multipliers (ADMM) [27] has attracted a great deal of attention in biostatistics. It mainly deals with convex optimization problems with constraints. The ADMM framework divides a problem into multiple subproblems that can be solved simultaneously.

From the viewpoint of tumor classification, there are commonly used methods for mammography classification, such as artificial neural networks [28, 29], nearest neighbor [30] and support vector machine (SVM) [31]. However, most of these methods rely on learning model parameters. Sparse representation-based classification (SRC) was originally proposed by Wright et al. [32] for face recognition. Recently, SRC and its improved methods have been used in image-based tumor classification [33, 34]. It is worth noting that the success of SRC depends on enough training data of the same category. For tumor classification, however, it is difficult to acquire sufficient and effective unlabeled samples. On the other hand, the discrimination ability of SRC will be reduced when there is a small disturbance on representation error [35]. Our previous work [36] proposed an inverse projection-based pseudo-full-space representation classification (PFSRC) method and successfully used it for robust face recognition. PFSRC focused on exploiting complementary information between training samples and test samples by utilizing existing available face images. Our another previous work [21] proposed an inverse space sparse representation (ISSR) model for microarray gene expression data-based tumor classification.

Motivated by these works, a mammogram-based breast tumor classification scheme is proposed in this paper. The main contributions are as follows: (1) An LPML-LRNMF-based feature learning method is proposed by effectively combining complementary strengths from NMF and deep learning. (2) The LPML-LRNMF model is optimized by ADMM, and the corresponding convergence is analyzed. (3) The ISSR model is firstly used for mammogram-based breast tumor classification.

The remainder of this paper is organized as follows: Section 2 describes the methodology of the presented breast tumor classification, which consists of LPML-LRNMF model-based feature representation learning and the ISSR-based classification. Experiments and discussions are shown in Sect. 3. Finally, conclusions are discussed in Sect. 4.

2 Methodology

2.1 Layer-Wise Pre-training Multilayer Low-Rank NMF Model

In this subsection, an improved NMF method, LPML-LRNMF, is proposed and used for feature representation learning of mammograms.

Lee and Seung [12] proposed an NMF based on multivariate analysis and linear algebra. Suppose $V \in \mathbb{R}^{p\times q}$ is a nonnegative matrix, which is decomposed into nonnegative basis matrix $W \in \mathbb{R}^{p\times r}$ and coefficient matrix $H \in \mathbb{R}^{r\times q}$.

$$V \approx WH.$$

It is worth noting that V is a collection of training samples and test samples in this paper. The conventional approach to find W and H is to minimize the error between V and WH [37], and the object function to be optimized is as follows:

$$\min_{W\geqslant 0, H\geqslant 0} \sum_{i=1}^{p}\sum_{j=1}^{q}(V_{ij}-(WH)_{ij})^2 = \min_{W\geqslant 0, H\geqslant 0}\|V-WH\|_F^2, \tag{2.1}$$

where $\|\cdot\|_F$ is the Frobenius norm and V_{ij}, $i = 1, \cdots p,\ j = 1, \cdots, q$, represent the elements in matrix V. Each column of H is an encoding correspondence with V. The rank r of the factorization is generally chosen so that $(p + q)r < p \times q$.

The original NMF, however, doesn't consider data characteristics or actual problem requirements into the model and doesn't fully dig the useful information hidden in feature matrix H. A priori information based on the characteristics of the data can be added as a regularization constraint of the model.

The proposed feature representation learning model aims to deal with the following three issues: (1) layer-wise pre-training strategy is introduced to mitigate the effect of the initial value on the NMF model; (2) low-rank constraint is added into the model based on intrinsic characteristic of mammogram data; (3) multilayer decomposition is performed to further mine deep representation feature information hidden in data.

The objective function of the LPML-LRNMF can be written as follows:

$$\min_{W_1\geqslant 0, H_1\geqslant 0}\|V-W_1H_1\|_F^2+\|H_1\|_*, \tag{2.2a}$$

$$\min_{W_2\geqslant 0, H_2\geqslant 0}\|H_1-W_2H_2\|_F^2+\|H_2\|_*,$$

$$\vdots \tag{2.2b}$$

The model (2) is based on the fact that the optimal output of the former layer is as the input of the latter layer, and so on. Suppose the decomposition level is L, model (2.2) can be simplified as the following form:

$$\min_{W_l\geqslant 0, H_l\geqslant 0}\|H_{l-1}-W_lH_l\|_F^2+\|H_l\|_*,\ l=1,\cdots,L, \tag{2.3}$$

where the initial matrix H_0 represents V, $W_l \in \mathbb{R}^{r_{l-1}\times r_l}$ and $H_l \in \mathbb{R}^{r_l\times q}$ represent the corresponding basis matrices and coefficient matrices of each layer, respectively. r_l $(l = 1, \cdots, L)$ represent the matrix decomposition dimensions, r_0 represents p, and $r_l \ll \min\{r_{l-1}, q\}$. Equations (2.2) and (2.3) are called the LPML-LRNMF model.

2.2 Optimization of LPML-LRNMF Model by ADMM

In this subsection, the LPML-LRNMF model is optimized by ADMM. From the optimization point of view, each layer of the model is similar, where the layer-wise pre-training technique means that the obtainable optimal solution of the previous layer is regarded as the input of the latter layer.

The optimization process for the each layer of LPML-LRNMF [Eq. (2.3)] can be rewritten as

$$\begin{aligned} &\min_{W_l \geqslant 0,\, H_l \geqslant 0} \|H_{l-1} - W_l H_l\|_F^2 + \|\mathcal{Z}_l\|_* \\ &\text{s.t. } \mathcal{Z}_l - H_l = 0, \end{aligned} \tag{2.4}$$

where $l = 1, \cdots, L$.

Let $\mathcal{J}(W_l, H_l, \mathcal{Z}_l) = \|H_{l-1} - W_l H_l\|_F^2 + \|\mathcal{Z}_l\|_*$, the augmented Lagrangian function of the problem (2.4) is defined by

$$\mathcal{L}(W_l, H_l, \mathcal{Z}_l) = \mathcal{J}(W_l, H_l, \mathcal{Z}_l) + \langle \Phi_l, \mathcal{Z}_l - H_l \rangle + \frac{\sigma}{2}\|\mathcal{Z}_l - H_l\|_F^2, \tag{2.5}$$

where $\sigma > 0$ is the penalty parameter, $\Phi_l \in \mathbb{R}^{r_l \times n}$ is the Lagrange multiplier and $\langle ., . \rangle$ is the inner product.

The ADMM scheme of Eq. (2.5) takes the following iteration:

$$W_l^{k+1} = \arg\min_{W_l \geqslant 0} \mathcal{L}(W_l, H_l^k, \mathcal{Z}_l^k), \tag{2.6a}$$

$$H_l^{k+1} = \arg\min_{H_l \geqslant 0} \mathcal{L}(W_l^{k+1}, H_l, \mathcal{Z}_l^k), \tag{2.6b}$$

$$Z_l^{k+1} = \arg\min_{Z_l \geqslant 0} \mathcal{L}(W_l^{k+1}, H_l^{k+1}, \mathcal{Z}_l), \tag{2.6c}$$

$$\Phi_l^{k+1} = \Phi_l^k + \sigma(\mathcal{Z}_l^{k+1} - H_l^{k+1}). \tag{2.6d}$$

Firstly, W_l is optimized for a given H_l. The subproblem W_l^{k+1} can be approximated by Eq. (2.6a):

$$W_l^{k+1} = \arg\min_{W_l \geqslant 0} \left\| H_{l-1} - W_l H_l^k \right\|_F^2.$$

Since the objective function (2.6a) is quadratic with respect to W_l, and the feasible region $W_l \geqslant 0$ is convex, we can guarantee that there exists local minimum.

Similar to [38], an iterative update rule is given as follows:

$$W_l^{k+1} = W_l^k - \mu(W_l^k H_l^k - H_{l-1})(H_l^k)^{\mathrm{T}}, \tag{2.7}$$

where $\mu > 0$ is iteration step. In order to satisfy the nonnegative constraints $W_l \geqslant 0$, projection operators can be constructed as follows, and any negative values in W_l^{k+1} are set to zero:

$$(\hat{w}_l^{k+1})_{ij} = \begin{cases} (w_l^{k+1})_{ij}, & \text{if } (w_l^{k+1})_{ij} \geqslant 0, \\ 0, & \text{otherwise,} \end{cases} \quad (w_l^{k+1})_{ij} \in W_l^{k+1}, \tag{2.8}$$

where $(w_l^{k+1})_{ij}$ are elements in W_l^{k+1}.

Subproblem H_l^{k+1} can be approximated by Eq. (2.6b):

$$\begin{aligned} H_l^{k+1} &= \arg\min_{H_l \geqslant 0} \left\| H_{l-1} - W_l^{k+1} H_l \right\|_F^2 + \left\langle \Phi_l^k, \mathcal{Z}_l^k - H_l \right\rangle + \frac{\sigma}{2} \left\| \mathcal{Z}_l^k - H_l \right\|_F^2 \\ &= \arg\min_{H_l \geqslant 0} \left\| H_{l-1} - W_l^{k+1} H_l \right\|_F^2 + \frac{\sigma}{2} \left\| \mathcal{Z}_l^k - H_l + \frac{\Phi_l^k}{\sigma} \right\|_F^2. \end{aligned} \tag{2.9}$$

Similar to W_l^{k+1}, the feasible region $H_l \geqslant 0$ is convex, which guarantees that there exists local minimum.

Let

$$\mathcal{F}(H_l) = \left\| H_{l-1} - W_l^{k+1} H_l \right\|_F^2 + \frac{\sigma}{2} \left\| \mathcal{Z}_l^k - H_l + \frac{\Phi_l^k}{\sigma} \right\|_F^2,$$

then

$$H_l^{k+1} = \arg\min_{H_l \geqslant 0} \mathcal{F}(H_l).$$

Let $\frac{\mathrm{d}\mathcal{F}(H_l)}{\mathrm{d}H_l} = 0$, similar to [37, 38], an iterative update rule is given as follows:

$$H_l^{k+1} = H_l^k .* [(W_l^{k+1})^{\mathrm{T}} H_{l-1} + \sigma Z_l^k + \Phi_l^k] ./ [(W_l^{k+1})^{\mathrm{T}} W_l^{k+1} H_l^k + \sigma], \tag{2.10}$$

where $.*$ and $./$ denote element-wise multiplication and division, respectively. The subtraction of the scalar σ is done to every element of the matrix $(W_l^{k+1})^{\mathrm{T}} W_l^{k+1} H_l^k$. For the nonnegative constraints $H_l \geqslant 0$, similar to Eq. (2.8), the projection operators can be constructed as follows:

$$(\hat{h}_l^{k+1})_{ij} = \begin{cases} (h_l^{k+1})_{ij}, & \text{if } (h_l^{k+1})_{ij} \geqslant 0, \\ 0, & \text{otherwise,} \end{cases} \quad (h_l^{k+1})_{ij} \in H_l^{k+1}, \tag{2.11}$$

where $(h_l^{k+1})_{ij}$ are elements in H_l^{k+1}.

Subproblem $\mathcal{Z}_l^{k+1}$ can be approximated by Eq. (2.6c):

$$\mathcal{Z}_l^{k+1} = \operatorname*{argmin}_{\mathcal{Z}_l \geqslant 0} \|\mathcal{Z}_l\|_* + \left\langle \Phi_l^k, \mathcal{Z}_l - H_l^{k+1} \right\rangle + \frac{\sigma}{2} \left\| \mathcal{Z}_l - H_l^{k+1} \right\|_F^2$$

$$= \underset{\mathcal{Z}_l \geqslant 0}{\operatorname{argmin}} \|\mathcal{Z}_l\|_* + \frac{\sigma}{2}\left\|\mathcal{Z}_l - (H_l^{k+1} - \Phi_l^k/\sigma)\right\|_F^2$$
$$= D_{1/\sigma}\left(H_l^{k+1} - \Phi_l^k/\sigma\right), \tag{2.12}$$

where $D(\cdot)$ is the singular value threshold operator [39]. For $\mathcal{Z}_l \geqslant 0$, similar to Eqs. (2.8) and (2.11), the projection operators can be constructed as follows:

$$(\hat{z}_l^{k+1})_{ij} = \begin{cases} (z_l^{k+1})_{ij}, & \text{if } (z_l^{k+1})_{ij} \geqslant 0, \\ 0, & \text{otherwise,} \end{cases} \quad (z_l^{k+1})_{ij} \in \mathcal{Z}_l^{k+1}, \tag{2.13}$$

where $(z_l^{k+1})_{ij}$ are elements in $\mathcal{Z}_l^{k+1}$.

Stopping criterion: $\max\left\{\left\|W_l^{k+1} - W_l^k\right\|_2, \left\|H_l^{k+1} - H_l^k\right\|_2, \left\|Z_l^{k+1} - Z_l^k\right\|_2\right\} \leqslant \varepsilon_1$, and $\left\|Z_l^{k+1} - H_l^{k+1}\right\|_F^2 \leqslant \varepsilon_2$.

Algorithm 1 Optimization of LPML-LRNMF.

Input: A non-negative matrix V. Given $\varepsilon_1, \varepsilon_2 > 0$, and $\sigma > 0$, $r_l \ll \min\{r_{l-1}, q\}$, the stopping criterion is $k_{\max} = 500$.

Initialize W_l^0, H_l^0 as non-negative matrix, $k = 0$.

while stopping criteria not satisfied **do**

Step 1. Update the variable W_l^{k+1} according to Eq. (2.7) and Eq. (2.8);

Step 2. Update the variable H_l^{k+1} according to Eq. (2.10) and Eq. (2.11);

Step 3. Update the variable $\mathcal{Z}_l^{k+1}$ according to Eq. (2.12) and Eq. (2.13);

Step 4. Update the Lagrange multiplier according to Eq. (2.6d);

Step 5. $k := k+1$, and go on Step 1.

end while

Output: An optimal solution can be obtained.

2.3 Convergence Analysis

Convergence analysis is crucial to optimization. Please see "Appendix A" for the corresponding convergence lemmas and theorems, and refer to [26] for the detailed proof of Theorem A.1. In Sect. 3.2, experiments will further demonstrate the convergence.

2.4 Inverse Space Sparse Representation Classification Model

2.4.1 Inverse Space Representation

Suppose $X = [x_1, \cdots, x_{s_1}, \cdots, x_{s_c}] \in \mathbb{R}^{d \times s_c}$ is a training sample set, $X_j = [x_{s_{j-1}+1}, \cdots, x_{s_j}] \in \mathbb{R}^{\mathrm{d} \times (s_j - s_{j-1})}$ are the jth category samples, where $j = 1, \cdots, c$ is the index of category. $Y = [y_1, \cdots, y_k] \in \mathbb{R}^{d \times k}$ is a test sample set.

SRC [32] assumes that each test sample $y_t \in R^d, t = 1, \cdots, k$ can be linearly represented by the training samples from the same category:

$$y_t = \gamma_{1,1}x_1 + \cdots + \gamma_{t,i}x_i + \cdots + \gamma_{s_c,i}x_i = \sum_{i=1}^{s_c} \gamma_{t,i}x_i = X\gamma_t, \tag{2.14}$$

where $\gamma_t = [\gamma_{t,1}, \cdots, \gamma_{t,s_c}]^{\mathrm{T}}$ is the corresponding coefficient vector. Without causing confusion, the corresponding projection way and representation space of SRC are called positive projection and positive space. PFSRC [36], by contrast, represented each training sample x_i by its corresponding pseudo-full-space $V_i = \{X, Y\} - \{x_i\}$, $i = 1, \cdots, s_c$, where the projection way is inverse to SRC and called inverse projection. It is worth noting that the PFSRC aims to explore complementary information contained in available face samples. However, there is no such obvious complementarity between tumor image data, and there are few effective labeled patient samples. To tackle this problem, an inverse space representation is proposed in our previous work [21]. The inverse space representation means that a training sample x_i is represented by its corresponding test sample space Y.

$$x_i = \alpha_{i,1}y_1 + \cdots + \alpha_{i,t}y_t + \cdots + \alpha_{i,k}y_k = \sum_{t=1}^{k} \alpha_{i,t}y_t = Y\alpha_i, \tag{2.15}$$

where $\alpha_{i,t} \in \mathbb{R}$ are representation coefficients and $\alpha_i = [\alpha_{i,1}, \cdots \alpha_{i,t}, \cdots, \alpha_{i,k}]^{\mathrm{T}}$ represents coefficient vector. By comparing Eqs. (2.14) and (2.15), one can observe that the differences between standard sparse representation and inverse space representation are projection ways and representation spaces.

Comparing Figs. 1(a) and (b), it is easy to notice that inverse space representation addresses the column coefficients before the test samples, rather than the row coefficients of training samples for standard sparse representation. Different projection ways make the inverse space representation less sensitive to the number of training samples than that of standard sparse representation [36].

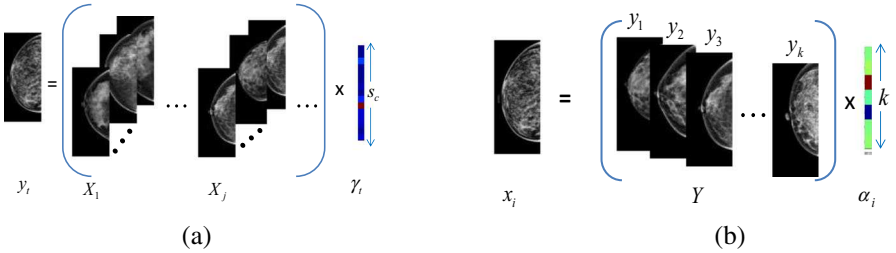

Fig. 1 Comparison of different representation ways: (a) standard sparse representation; (b) inverse space representation

Considering the sparsity between categories, the sparsity constraint can be introduced into the inverse space representation called the inverse space sparse representation (ISSR):

$$\min_{\alpha_i} \|x_i - Y\alpha_i\|_2^2 + \lambda\|\alpha_i\|_1, \tag{2.16}$$

where $\lambda > 0$ is regularization parameter and α_i is the representation coefficient vector of x_i.

2.4.2 Feasibility Analysis of ISSR Model

The feasibility of the ISSR model is verified similar to [32]; for the simplicity of analysis, the regular term in Eq. (2.16) is removed, and then, the representation becomes a least square problem:

$$\hat{\alpha}_i = \underset{\alpha_i}{\operatorname{argmin}} \|x_i - Y\alpha_i\|_2^2.$$

Let x_i^j represent a training sample that belongs to category j and can be represented by the test sample space. Suppose Y^j denotes test sample subspace that belongs to the same category with x_i, the associated representation $\hat{x}_i^j = \sum_j Y^j \delta_j(\hat{\gamma}_i)$ is actually the perpendicular projection of x_i onto the test sample full space Y. The reconstruction error by each category $e_j = ||x_i^j - Y^j\delta_j(\hat{\alpha}_i)||_2^2$ is used for classification. It can be readily derived by

$$e_j = ||x_i^j - Y^j\delta_j(\hat{\alpha}_i)||_2^2 = ||x_i^j - \hat{x}_i^j||_2^2 + ||\hat{x}_i^j - Y^j\delta_j(\hat{\alpha}_i)||_2^2.$$

Obviously, it is the amount $e_j^* = ||\hat{x}_i^j - Y^j\delta_j(\hat{\alpha}_i)||_2^2$ that works because $||x_i^j - \hat{x}_i^j||_2^2$ is a constant for all categories.

Denoted by $\chi_j = Y^j\delta_j(\hat{\alpha}_i)$ and $\hat{\chi}_j = \sum_{m\neq j} Y^m\delta_m(\hat{\alpha}_i)$, $m = 1, \cdots, c$, $m \neq j$, since $\hat{\chi}_j$ is parallel to $\hat{x}_i^j - Y^j\delta_j(\hat{\alpha}_i)$, one can readily have

$$\frac{||\hat{x}_i^j||_2}{\sin(\chi_j, \hat{\chi}_j)} = \frac{||\hat{x}_i^j - Y^j\delta_j(\hat{\alpha}_i)||_2}{\sin(\chi_j, \hat{x}_i^j)},$$

where $(\chi_j, \hat{\chi}_j)$ is the angle between χ_j and $\hat{\chi}_j$, and $(\chi_j, \hat{x}_i^j)$ is the angle between χ_j and $\hat{x}_i^j$.

So, the representation error can be represented by

$$e_j^* = ||\hat{x}_i^j - Y^j\delta_j(\hat{\alpha}_i)||_2^2 = \frac{\sin^2(\chi_j, \hat{x}_i^j)||\hat{x}_i^j||_2^2}{\sin^2(\chi_j, \hat{\chi}_j)}. \tag{2.17}$$

Equation (2.17) shows that the ISSR is effective and robust by a "double checking," because we need not only consider if sin $(\chi_j, \hat{x}_i^j)$ is small, but also consider if sin $(\chi_j, \hat{\chi}_j)$ is large. If x_i^j has a strong correlation with a test sample.

2.4.3 Stability Analysis of ISSR Model

Theorem (Classification Stability of ISSR) *Suppose x_i and x_j are the i*th *and j*th *training samples, and the relationship between x_i and x_j is $x_j = x_i + \Delta(x_i)$, where $\Delta(x_i)$ is a disturbance of x_i. Based on the test samples Y, the inverse space representations of x_i, x_j are as follows*: $x_i = Y\alpha_i$ and $x_j = Y\alpha_j$, *where α_i and α_j are representation coefficients, respectively. Let $\Delta(Y)$ represent the disturbance corresponding to $\Delta(x_i)$. If*

$$\varepsilon = \max\left\{\frac{\|\Delta(x_i)\|_2}{\|x_i\|_2}, \frac{\|\Delta(Y)\|_2}{\|Y\|_2}\right\} \leqslant \frac{\varphi_k(Y)}{\varphi_1(Y)},$$

and $\sin(\theta) = \rho_{LS}/\|x_i\|_2 \neq 1$, *where* $\rho_{LS} = ||Y\alpha_{LS_i} - x_i||_2$, $\alpha_{LS_i} = \arg\min_{\alpha_i} ||x_i - Y\alpha_i||_2$, *then*

$$\frac{||\alpha_j - \alpha_i||_2}{||\alpha_i||_2} \leqslant \varepsilon\left\{\frac{2\kappa_2(Y)}{\cos(\theta)} + \tan(\theta)\kappa_2(Y)^2\right\} + O(\varepsilon^2), \tag{2.18}$$

where $\kappa_2(Y)$ ($\kappa_2(Y) = ||Y||_2 \cdot ||(Y^{\mathrm{T}}Y)^{-1}Y^{\mathrm{T}}||_2$, $\kappa_2(Y)^2 = ||Y||_2^2 \cdot ||(Y^{\mathrm{T}}Y)^{-1}||_2$) is the l_2-norm conditional number of Y and θ is angle between x_i and its projection vector on Y.

The conclusion indicates that the distance between α_i and α_j is very small when x_i is similar to x_j (in other words, Y has a small disturbance $\Delta(Y)$). From Eq. (2.18), one can see that coefficients are more sensitive to a small disturbance Δ than that of reconstruction error because, for nonzero residual problems, it is the square of the condition number that measures the sensitivity of coefficients. Moreover, it is worth noting that we focus on the column coefficient vector $\alpha_{1,1}, \alpha_{1,2}, \cdots, \alpha_{s_c,1}$ before each test sample when we calculate the category contribution rate (CCR) similar to [36]. The difference lies in the representation coefficients α of different representation spaces. The larger the CCR is, the higher the correlation is. However, it has been demonstrated that disturbance will affect row coefficients rather than column coefficients. Moreover, the effect on column coefficients is a positive impact when CCRs of different categories are calculated.

Please see "Appendix B" for the detailed proof of the classification stability theorem. And the classification stability of inverse space representation is verified more stable than reconstruction error [36].

2.5 Breast Tumor Classification Based on LPML-LRNMF and ISSRC

Because the coefficient matrix obtained by the NMF feature representation learning has sparse characteristics [12]. LPML-LRNMF is an improved method of NMF,

and the obtained coefficient matrix also has sparse features. A mammogram-based tumor classification scheme is proposed by integrating LPML-LRNMF feature representation learning and ISSRC. It is worth noting that the $V = [X, Y] \in \mathbb{R}^{d\times(s_c+k)}$ is a collection of training samples and test samples in LPML-LRNMF model, where $q = s_c + k$. Equation (2.16) can be rewritten as Eq. (2.19) when the samples are replaced by the corresponding LPML-LRNMF features. Suppose the feature matrix H_l of LPML-LRNMF is divided into the training set H_l^{train} and the test set H_l^{test}. For any $h_l^{\text{train}} \in H_l^{\text{train}}$, ISSR represents h_l^{train} by H_l^{test} as follows:

$$\min_{\alpha} \left\| h_l^{\text{train}} - H_l^{\text{test}}\alpha \right\|_2^2 + \lambda\|\alpha\|_1, \tag{2.19}$$

where $H_l^{\text{train}} = [(h_l^{\text{train}})_1, \cdots, (h_l^{\text{train}})_{s_c}], i = 1, \cdots, s_c$.

Algorithm 2 Mammograms-based breast tumor classification algorithm

Input: Training sample set $X=[x_1,\cdots,x_{s_c}]$, training label set $L=[l_1,l_2\cdots,l_{s_c}]$ and test sample set $Y=[y_1,y_2,\cdots,y_k]$.

1) By Eqs. (2.7)-(2.13), $l=1$, the first layer of LPML-LRNMF-basedfeature representation learning is realized, and the corresponding local optimal characteristics H_1 is obtained.
2) Then the output H_1 from the first layer is imported to the second layer of LPML-LRNMF. Similar to the optimization process at the first layer, by Eqs. (2.7)-(2.13), $l=2$, the second layer of LPML-LRNMF feature representation learning is realized.
3) By Eq. (2.19), the LPML-LRNMF feature results are imported into the ISSR model, where $X=H_2^{\text{train}}$, $H_2^{\text{train}}=[(h_2^{\text{train}})_1,\cdots,(h_2^{\text{train}})_{s_c}]$. $Y=H_2^{\text{test}}$, $H_2^{\text{test}}=[(h_2^{\text{test}})_1,\cdots,(h_2^{\text{test}})_k]$.
4) The CCR matrix is obtained, and by normalizing the CCR matrix, relationships between each test sample and all categories are obtained.

Output: Each test sample is classified into the category with the maximal CCR.

Algorithm 2 is the mammogram-based breast tumor classification algorithm based on LPML-LRNMF-based feature representation learning and ISSRC. The corresponding flowchart is shown in Fig. 2.

3 Experiments and Discussions

In this subsection, the performance of the proposed method is demonstrated on the three aspects: (1) the convergence of LPML-LRNMF by ADMM optimization is verified, (2) the feature representation learning performance of LPML-LRNMF is tested and (3) the classification performance of ISSRC is compared with classical classification methods and the state-of-the-art results. Without loss of generality, the tenfold cross-validation and two-layer LPML-LRNMF for feature learning are used in all experiments. Experiments are carried out using MATLAB R2016a and R-3.4.1 on a

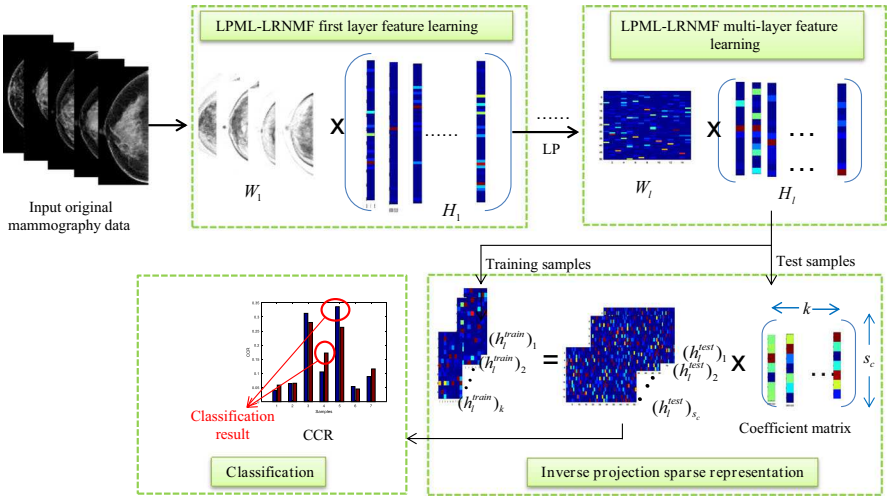

Fig. 2 Proposed mammogram-based breast tumor classification framework

3.30-GHz machine with 4.00 GB RAM. It is noted that LPML-LRNMF-based feature representation learning scheme can be done multilayer. In fact, some information will be lost as the number of decomposition layers increases. Therefore, it is not that the more the layers are decomposed, the better the feature classification effect will be. We have done experiments on Zhejiang Cancer Hospital (ZCH) dataset to select the optimal decomposition level. The classification accuracies from one layer to four layers are 72.86%, 90.89%, 85.47% and 74.42%, which show that two layers can achieve good result in our work. Hence, the subsequent experiments are all based on two layers.

3.1 Breast Tumor Datasets

Experiments have conducted on two datasets: one is the public test dataset provided by the Mammographic Image Analysis Society (MIAS) [40] (http://peipa.essex.ac.uk/info/mias.html), and the other is the actual clinical dataset provided by Zhejiang Cancer Hospital (ZCH). The MIAS dataset has 322 mammograms with size of 1 024 × 1 024 from 161 samples. ZCH dataset has 688 mammograms with size of 3 328 × 2 560 from 172 samples. Without loss of generality, we randomly select 104 normal samples, 31 benign samples and 26 malignant samples from the MIAS dataset and 68 normal samples and 17 malignant samples from the ZCH dataset. For the convenience of experiment and calculation, these samples have been cut out the non-breast background area and adjusted to 512 × 306 and 303 × 128, respectively. Some examples are shown in Fig. 3.

3.2 Convergence Analysis of LPML-LRNMF

In Sect. 2.2, optimization of LPML-LRNMF model by ADMM. Here, the corresponding convergence is analyzed. The results are shown in Fig. 4, Fig. 4(a) shows the

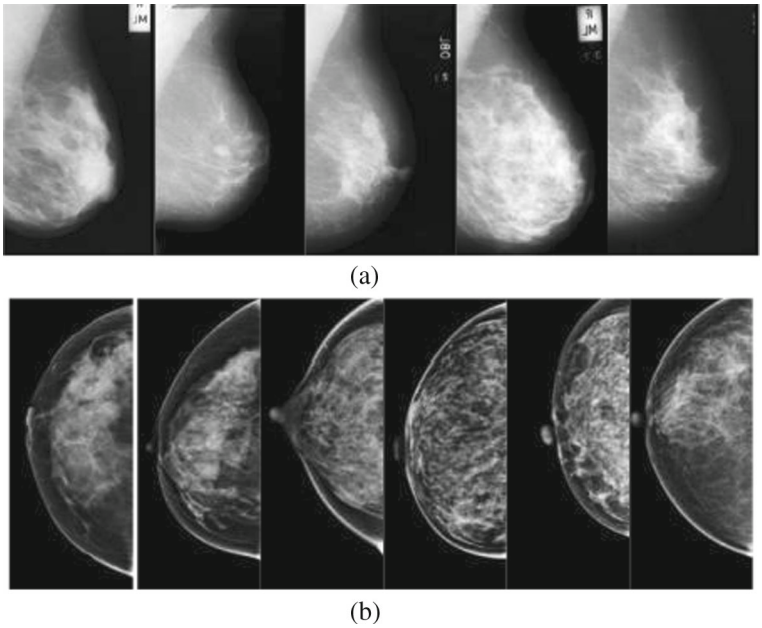

Fig. 3 Some examples of mammogram images: (a) MIAS dataset, (b) ZCH dataset

iteration error between exact and iterative solutions, and Fig. 4(b) shows the iteration error between the adjacent iterations. And Fig. 4(c) shows the trend graph, which shows that the solution gradually becomes stable and converges to the numerical solution. It can be seen from Fig. 4(a) that the convergence error between exact and iterative solutions of ADMM is about 1e−5, and iteration time is about 50 s. Figure 4(b) shows that the convergence error between the adjacent iterations is about 0.02, and iteration time is about 100 s. The experiment verifies that ADMM optimization achieves good convergence of LPML-LRNMF model.

3.3 Representation Performance of LPML-LRNMF Model

In this subsection, the effectiveness and efficiency of the proposed feature representation learning method, LPML-LRNMF, are demonstrated by analyzing mean, variance, feature expression level line chart and entropy. Without causing confusion, V represents the original image matrix, H_1 and H_2 represent the first- and second-layer feature matrix of LPML-LRNMF. The decomposition dimensions corresponding to the first and second layers are $r_1 = 50$ and $r_2 = 15$ by experience.

In order to verify the feature representation performance of LPML-LRNMF, the correlation analysis is done between normals and the mean correlation coefficient of all 17 malignants. Figure 5 shows the correlation coefficients of normals and the mean sample. In Fig. 5, blue line and red line correspond to original data and multilayer feature, respectively. It can be seen that the correlation coefficients of normals and the mean sample of malignants are generally smaller than those of normals. It is also can be

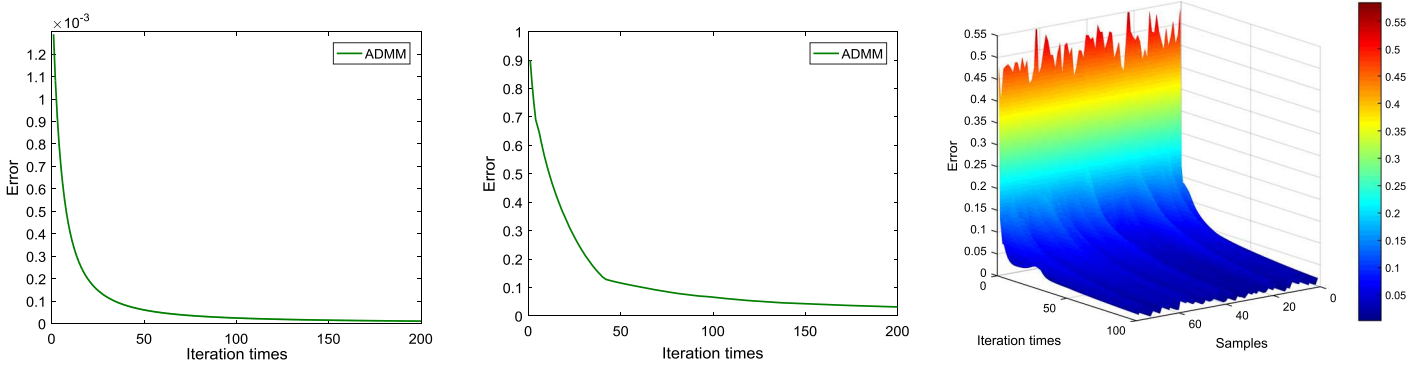

Fig. 4 Convergence analysis of LPML-LRNMF by ADMM optimization: (a) iteration error between the exact and iterative solutions, (b) iteration error between the adjacent iterations, (c) optimization solution trend graph

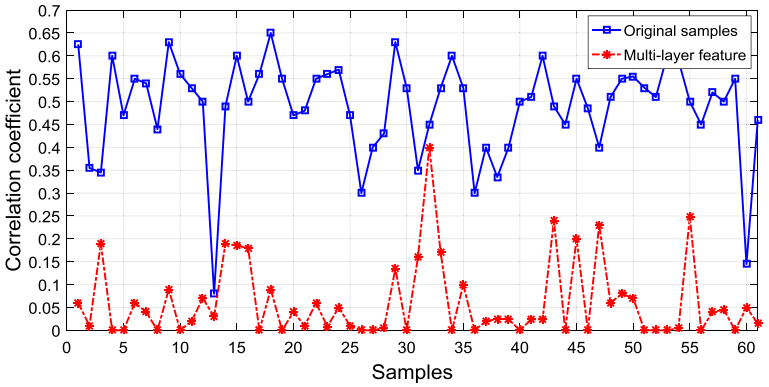

Fig. 5 Correlation coefficients of normals and the mean sample of malignants on ZCH dataset. (Color figure online)

observed that the correlation coefficients between normals and malignants gradually decrease as the decomposition layers increase.

The expression profiles of features for the normals and malignants are analyzed in Fig. 6. Figure 6(a) represents the original matrix V, and Figs. 6(b) and (c) show the first- and second-layer hidden components matrices H_1 and H_2. In Fig. 6, the red curves denote the feature expression levels of the normals and the blue curves express those of the malignants. The horizontal straight lines indicate the mean values of feature expression levels in the corresponding category. In the case of H_2, the difference in the mean values is large. For V, the difference in the mean values is basically 0. Moreover, considerable fluctuation can be seen between the binary category and the irrelevant features in terms of standard deviation (std). It is implied that feature obtained by LPML-LRNMF is easier to distinguish normals and malignants than original data.

For further verifying the classification performance of LPML-LRNMF, the entropy analysis is done. Entropy is a measure of uncertainty. The smaller the entropy is, the lower the uncertainty of representation is, and the better the feature is. In Fig. 7, blue line, green line and red line correspond to original data V, the first-layer feature H_1 and the second-layer feature H_2, respectively. One can observe that the entropy of all samples gradually declines as the decomposition layers increase. This implied that the features obtained by LPML-LRNMF are more conductive to classification.

3.4 Classification Performance of LPML-LRNMF Model

For further accessing the classification performance, experiments are conducted on ZCH and MIAS datasets from aspects. Firstly, the classification performance of the LPML-LRNMF is compared with different feature representation learning methods. And then, the classification performance of ISSRC is compared with different classifiers. Finally, our finally classification result is compared with the latest published classification results.

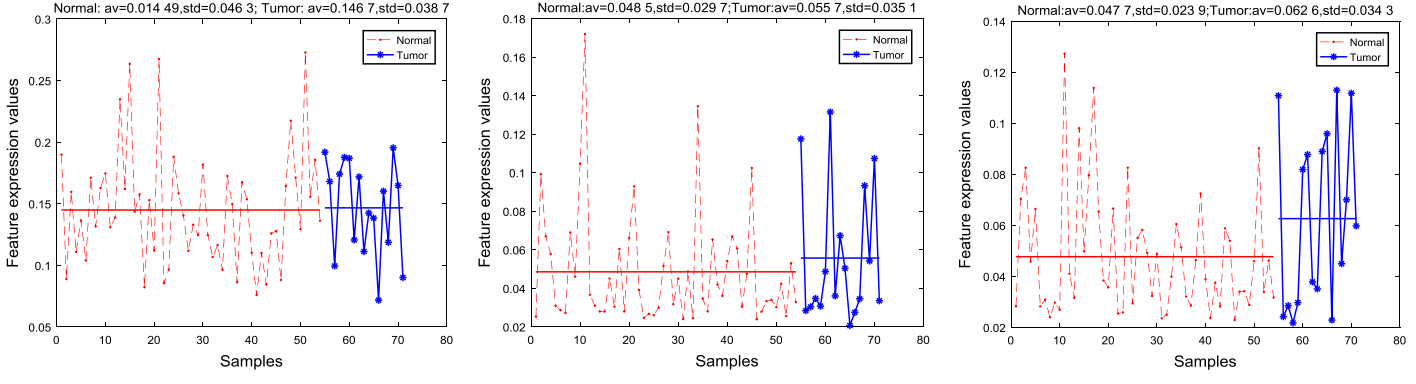

Fig. 6 Comparison of feature expression levels for normals and malignants on ZCH dataset. (a) V, (b) H_1 and (c) H_2 are original gene data, the first- and the second-layer feature matrix. (Color figure online)

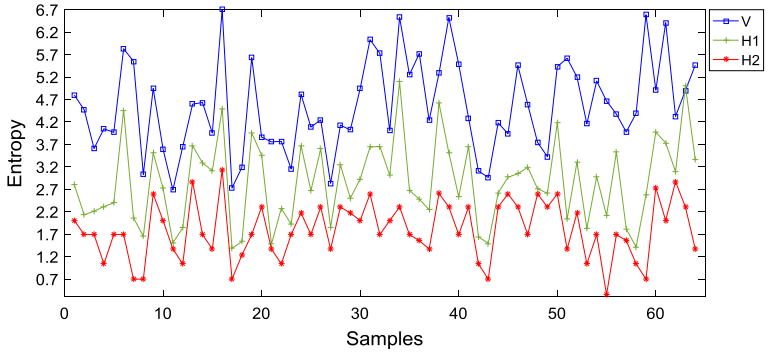

Fig. 7 Comparison of entropy for original image (V) and different layer features (H_1, H_2) on ZCH dataset. (Color figure online)

Table 1 Classification performance of different feature learning methods on ZCH dataset

Methods	Accuracy /%	Sensitivity /%	Specificity /%	AUC	Time /s
Original data	45.89	55.00	54.05	0.535 2	–
NMF [12]	68.04	75.00	69.05	0.778 2	120.9
LRNMF [20]	72.86	75.00	71.67	0.772 4	120.3
LPML-SNMF [21]	85.89	90.00	85.00	0.891 0	128.5
Our method	**90.89**	**96.67**	**91.67**	**0.952 8**	**126.3**

3.4.1 Comparison of Different Feature Learning Methods

In this subsection, different feature representation learning methods, NMF [12], LRNMF [20] LPML-SNMF [21], our method (LPML-LRNMF) are adopted, and the same classification method ISSRC is adopted. Table 1 gives the results of accuracy, sensitivity, specificity, AUC and running times. Experimental results show that LPML-LRNMF has a comparable running time and a much higher accuracy, sensitivity, specificity and AUC than other methods.

In order to give a more intuitive comparison of different methods, box plots of error rates, receive operating characteristic curve (ROCC) [41] analysis, precision recall curve (PRC) [42] analysis and decision curve analysis (DCA) [43] are shown in Fig. 8. Red line in Fig. 8(a) shows the average error rate of original data, Lee [12], Li [20], Yang [21], and our method. The result shows that LPML-LRNMF-based ISSRC has the lowest error rate and the best classification performance. In Fig. 8(b), the larger the area under the higher the ROCC is, the better the model is. In Fig. 8(c), the higher the DCA is, the better the model is. DCA is a way of evaluating models by maximizing the clinic net benefit (NB) of profit minuses harm. In other words, the higher the DCA

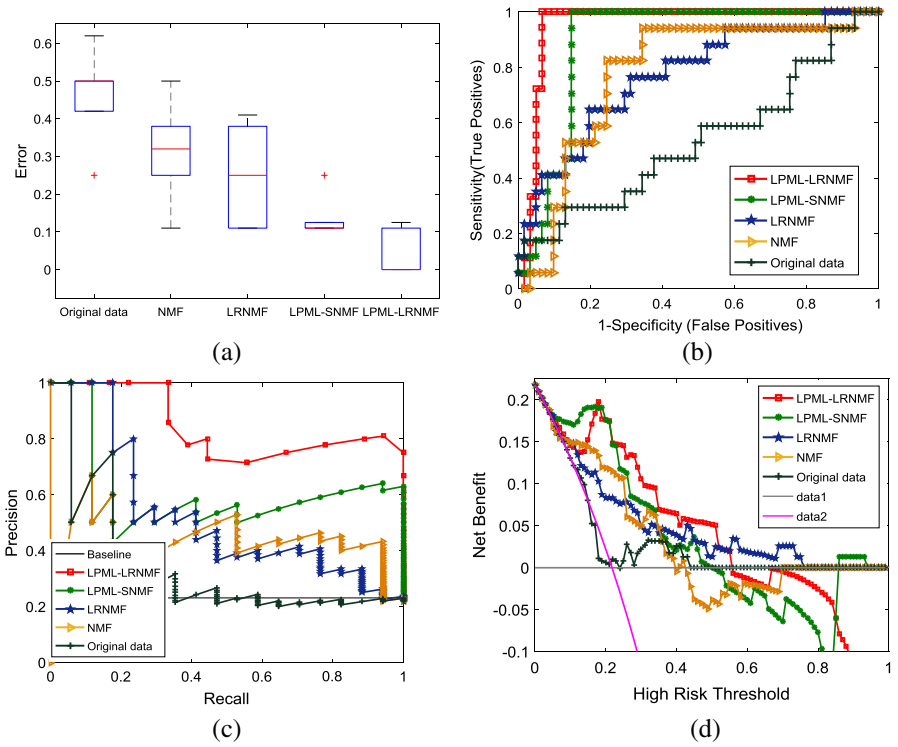

Fig. 8 Comparison of different methods: (a) box plots for error rates, (b) ROCC analysis, (c) PRC analysis, (d) DCA analysis. (Color figure online)

is, the smaller the loss of the model is. In Fig. 8(d), the closer the PRC curve is to the upper right corner is, the better the model is. As can be seen from Fig. 8, all the four indexes show that LPML-LRNMF-based ISSRC is better than other methods.

3.4.2 Comparison of Different Classification Methods

In this subsection, the classification performance of ISSRC is compared with some classic and updated classifiers, such as NN [30], SVM [31] and SRC [32], PFSRC [36], RRC_L1 and RRC_L2 [44]. Among them, RRC_L1 and RRC_L2 are recently proposed SRC-based RRC coding models with L_1 and L_2 constraints, respectively, PFSRC is another improved SRC method for face recognition. In Sect. 3.4.1, the advantages of the LPML-LRNMF-based feature learning have been verified, so all classifiers are based on LPML-LRNMF in this subsection.

In addition to the commonly used classification accuracy, the error reduction rate [45] $\text{ERR} = (\text{ER}_1 - \text{ER}_2)/\text{ER}_1 \times 100\%$ is also adopted, where ER_1 is the error rate of the other classifiers classification result, ER_2 is the error rate of our method classification result, and ERR is denoted by a notion $\downarrow$.

Table 2 Classification performance of different classifiers on ZCH dataset

Methods	Error rates /%	ERR /%	Time /s
ISSRC	**9.11**	–	126.3
PFSRC [36]	15.36	↓40.69	**124.5**
RRC_L2 [44]	15.36	↓40.69	128.2
RRC_L1 [44]	17.86	↓48.99	149.8
SRC [32]	16.61	↓45.15	135.9
SVM [31]	19.29	↓52.77	167.1
NN [30]	20.89	↓56.39	154.7

Table 3 Classification performance with the latest published results on MIAS dataset

Experiments	Methods	Accuracy /%
Our method	**LPML-LRNMF-based ISSRC**	**95.45**
Setiawan (2015) [46]	LAWS-ANN	93.90
Kutluk (2013) [47]	LVQ	90.00
Rampun (2018) [48]	LQP	85.60
Herwanto (2013) [49]	CPAR	83.00

Table 2 gives the comparison of ERRs and classification times. From Table 2, one can notice that ISSRC is superior to classical classifiers and updated classifiers, which embodies in the lowest classification error rates and almost the same running time.

3.4.3 Comparison with State-of-the-Art Results

In this subsection, the proposed method is compared with the latest published classification results on public test MIAS dataset. Table 3 shows that the classification accuracy rate of our method is 95.45%, which is the highest than those of latest published results given in [46–49].

4 Conclusions

In this paper, a mammogram-based breast tumor classification scheme is proposed by integrating LPML-LRNMF feature representation learning and ISSRC. The LPML-LRNMF is constructed by integrating hierarchical learning, layer-wise pre-training strategy and considering the low-rank characteristics of mammogram image as a prior constraint. Moreover, the model is solved by ADMM and the corresponding convergence analysis is given. The classification is fulfilled by ISSRC model, which is firstly used to exploit information embedded in the existing tumor mammogram images. Experiments on both public test mammogram image dataset from MIAS and the actual clinical dataset from ZCH show that the proposed breast tumor classification framework is superior to other compared methods.

There remain some interesting problems. One is how to further improve the feature representation learning model by mining and integrating intrinsic characteristics of multimodal breast data. The other is how to further optimize the model, such as adding more targeted prior information as regular terms and considering mixed driven of unlabeled data and model.

Acknowledgements The authors would like to thank Mammographic Image Analysis Society, London, UK, and Zhejiang Cancer Hospital for their breast datasets.

Appendix A

$$p^* = \inf\{\|H_{l-1} - W_l H_l\|_F^2 + \|\mathcal{Z}_l\|_* | \mathcal{Z}_l - H_l = 0, W_l \geqslant 0, H_l \geqslant 0\}. \tag{A.1}$$

Assumption A.1 The augmented Lagrangian $\mathcal{L}(W_l, H_l, \mathcal{Z}_l)$ has a saddle point. Explicitly, there exists $(W_l^*, H_l^*, \mathcal{Z}_l^*, \Phi_l^*)$, not necessarily unique, for which

$$L_0(W_l^*, H_l^*, \mathcal{Z}_l^*, \Phi_l) \leqslant L_0(W_l^*, H_l^*, \mathcal{Z}_l^*, \Phi_l^*) \leqslant L_0(W_l, H_l, \mathcal{Z}_l, \Phi_l^*)$$

holds for all $W_l, H_l, \mathcal{Z}_l, \Phi_l$. Note that L_0 ($\sigma = 0$) is the standard Lagrangian for the problem. By Assumption A.1, it follows that $L_0(W_l^*, H_l^*, \mathcal{Z}_l^*, \Phi_l^*)$ is finite for any saddle point $(W_l^*, H_l^*, \mathcal{Z}_l^*, \Phi_l^*)$. This implies that $(W_l^*, H_l^*, \mathcal{Z}_l^*)$ is a solution to the problem (2.4), so $\mathcal{Z}_l^* - H_l^* = 0$ and $\left\|H_{l-1} - W_l^* H_l^*\right\|_F^2 + \left\|\mathcal{Z}_l^*\right\|_* < \infty$.

Theorem A.1 *Suppose there exists the saddle point* $(W_l^*, H_l^*, \mathcal{Z}_l^*, \Phi_l^*)$, *and satisfying the* Assumption A.1, *let* $\{(W_l^k, H_l^k, \mathcal{Z}_l^k, \Phi_l^k)\}$ *be the sequence generated by Eq.* (2.6), *and then,*

(1) *Residual convergence*: $r^k \to 0$ *as* $k \to \infty$, *where* $r^k = \mathcal{Z}_l^k - H_l^k$, *i.e., the iterates approach feasibility*;
(2) *Objective convergence*: $\left\|H_{l-1} - W_l^* H_l^*\right\|_F^2 + \left\|\mathcal{Z}_l^*\right\|_* \to p^*$ *as* $k \to \infty$, *i.e., the objective function of the iterates approaches the optimal value*;
(3) *Dual variable convergence*: $\Phi_l^k \to \Phi_l^*$ *as* $k \to \infty$, *i.e., where* Φ_l^* *is a dual optimal point.*

Appendix B

Proof In order to discuss the value of $\frac{||\alpha_j - \alpha_i||_2}{||\alpha_i||_2}$, we need to find the relationship between α_i and α_j. Let $\alpha_i(t)$ be continuously differentiable for all $t \in [0, \varepsilon]$, where $\alpha_i = \alpha_i(0)$ and $\alpha_j = \alpha_i(\varepsilon)$. Let $\alpha_i(t)$ do the Taylor expansion at $t = 0$:$\alpha_i(t) = \alpha_i(0) + \varepsilon\alpha_i'(0) + O(t^2)$. We have $\alpha_j = \alpha_i + \varepsilon\alpha_i'(0) + O(\varepsilon^2)$ when $t = \varepsilon$. Then,

$$\frac{||\alpha_j - \alpha_i||_2}{||\alpha_i||_2} = \varepsilon \frac{||\alpha_i'(0)||_2}{||\alpha_i||_2} + O\left(\varepsilon^2\right). \tag{B.1}$$

In order to obtain $||\alpha_i'(0)||_2$, similar to Theorem 5.3.1 in [50], one can construct $(Y+tf)^{\mathrm{T}}(Y+tf)\alpha_i(t)$, where $f=\Delta(Y)/\varepsilon$; then,

$$(Y+tf)^{\mathrm{T}}(Y+tf)\alpha_i(t)=(Y+tf)^{\mathrm{T}}(x_i+t\Delta(Y)\alpha_i(t)/\varepsilon).$$

Let $E=\Delta(x_i)/\varepsilon$, then

$$(Y+tf)^{\mathrm{T}}(Y+tf)\alpha_i(t)=(Y+tf)^{\mathrm{T}}(x_i+tE). \tag{B.2}$$

In order to bound $||\alpha_i'(0)||_2$, one can take the derivative of Eq. (B.2) and set x_j, $f^{\mathrm{T}}Y\alpha_i+Y^{\mathrm{T}}f\alpha_i+YY^{\mathrm{T}}\alpha_i'(0)=Y^{\mathrm{T}}E+f^{\mathrm{T}}x_i$, i.e.,

$$\alpha_i'(0)=\left(Y^{\mathrm{T}}Y\right)^{-1}Y^{\mathrm{T}}(E-f\alpha_i)+\left(Y^{\mathrm{T}}Y\right)^{-1}f^{\mathrm{T}}(x_i-Y\alpha_i). \tag{B.3}$$

By singular value decomposition theorem [50], we have $\mathrm{rank}(Y+tf)=k$ for all $t\in[0,\varepsilon]$, where $\|\Delta(Y)\|_2\leqslant\varphi_k(Y)(\varphi_k(Y)$ is the largest singular value of $Y)$. Then,

$$||f||_2=||\Delta(Y)/\varepsilon||_2\leqslant\varphi_k(Y)\leqslant||Y||_2,$$

and $||E||_2=||\Delta(x_i)/\varepsilon||_2\leqslant||x_i||_2$.

By substituting Eq. (B.3) result into Eq. (B.1), taking norms, the inequality can be obtained:

$$\frac{||\alpha_j-\alpha_i||_2}{||\alpha_i||_2}\leqslant\varepsilon\left\{\|Y\|_2\cdot||(Y^{\mathrm{T}}Y)^{-1}Y^{\mathrm{T}}||_2\cdot\left(\frac{||x_i||_2}{||Y||_2||\alpha_i||_2}+1\right)+\frac{\rho_{LS}}{||Y||_2||\alpha_i||_2}\cdot\|Y\|_2^2\cdot||(Y^{\mathrm{T}}Y)^{-1}||_2\right\}+O\left(\varepsilon^2\right).$$

Since $Y^{\mathrm{T}}(Y\alpha_i-x_i)=0$, $Y\alpha_i$ is orthogonal to $Y\alpha_i-x_i$, it is also known that $\|x_i-Y\alpha_i\|_2^2+\|Y\alpha_i\|_2^2=\|x_i\|_2^2$, and then, $||Y||_2^2\cdot||\alpha_i||_2^2\geqslant||x_i||_2^2-\rho_{LS}^2$.

The relationship between α_i and α_j will be

$$\frac{||\alpha_j-\alpha_i||_2}{||\alpha_i||_2}\leqslant\varepsilon\left\{\kappa_2(Y)\left(\frac{1}{\cos(\theta)}+1\right)+\kappa_2(Y)^2\frac{\sin(\theta)}{\cos(\theta)}\right\}+O(\varepsilon^2).$$

References

[1] Gao, Y., Church, P.G.: Improving molecular cancer class discovery through sparse non-negative matrix factorization. Bioinformatics **21**, 3970–3975 (2005)

[2] Lou, P., Qian, W., Romilly, P.: CAD-aided mammogram training. Acad. Radiol. **12**, 1039–1048 (2005)

[3] Dorsi, C.J., Kopans, D.B.: Mammography interpretation: the BI-RADS method. Am. Fam. Phys. **55**, 1548–1550 (1997)

[4] Liu, S., Babbs, C.F., Delp, E.J.: Multiresolution detection of spiculated lesions in digital mammograms. IEEE Trans. Image Process. **10**, 874–884 (2001)

[5] Ebrahim, A.Y.: Detection of breast cancer in mammograms through a new features and decision tree based classification framework. J. Theor. Appl. Inf. Technol. **95**, 6256–6267 (2017)

[6] Catanzariti, E., Ciminello, M., Prevete, R.: Computer aided detection of clustered microcalcifications in digitized mammograms using Gabor functions. In: International Conference on Image Analysis and Processing, pp. 266–270 (2003)
[7] Oliver, A., Torrent, A., Llado, X., Marti, J.: Automatic diagnosis of masses by using level set segmentation and shape description. In: International Conference on Pattern Recognition, pp. 2528–2531 (2010)
[8] Rashed, E., Ismail, I., Zaki, S.: Multiresolution mammogram analysis in multilevel decomposition. Pattern Recognit. Lett. **28**, 286–292 (2007)
[9] Bengio, Y., Courville, A., Vincent, P.: Representation learning: a review and new perspectives. IEEE Trans. Pattern Anal. Mach. Intell. **35**, 1798–1828 (2012)
[10] Lecun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**, 436 (2015)
[11] Han, Z., Wei, B., Zheng, Y., Yin, Y., Li, K., Li, S.: Breast cancer multi-classification from histopathological images with structured deep learning. Sci. Rep. **7**, 4172 (2017)
[12] Lee, D., Seung, H.: Learning the parts of objects by non-negative matrix factorization. Nature **401**, 788–791 (1999)
[13] Sauwen, N., Sima, D., Acou, M., Achten, E., Maes, F.: A semi-automated segmentation framework for MRI based brain tumor segmentation using regularized nonnegative matrix factorization. In: International Conference on Signal-Image Technology and Internet-Based Systems, pp. 88–95 (2017)
[14] Tsinos, C.G., Rontogiannis, A., Berberidis, K.: Distributed blind hyperspectral unmixing via joint sparsity and low-rank constrained non-negative matrix factorization. IEEE Trans. Comput. Imaging **3**, 160–174 (2017)
[15] Liu, W., Peng, F., Feng, S., You, J., Chen, Z.: Semantic feature extraction for brain CT image clustering using nonnegative matrix factorization. In: Medical Biometrics, First International Conference, vol. 4901, pp. 41–48 (2008)
[16] Zheng, C.H., Ng, T.Y., Zhang, L., Shiu, C.K., Wang, H.Q.: Tumor classification based on non-negative matrix factorization using gene expression data. IEEE Trans. Nanobiosci. **10**, 86–93 (2011)
[17] Shang, R., Wang, W., Stolkin, R., Jiao, L.: Nonnegative spectral learning and sparse regression-based dual-graph regularized feature selection. IEEE Trans. Cybern. **48**, 793–806 (2017)
[18] Shang, R., Zhang, Z., Jiao, L., Wang, W., Yang, S.: Global discriminative-based nonnegative spectral clustering. Pattern Recognit. **55**, 172–182 (2016)
[19] Shang, R., Yuan, Y., Jiao, L., Hou, B., Esfahani, A.M.G.: A fast algorithm for SAR image segmentation based on key pixels. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. **99**, 1–17 (2017)
[20] Li, X., Cui, G., Dong, Y.: Graph regularized non-negative low-rank matrix factorization for image clustering. IEEE Trans. Cybern. **47**, 3840–3853 (2017)
[21] Yang, X.H., Wu, W., Chen, Y., Li, X., Zhang, J., Long, D., Yang, L.: An integrated inverse space sparse representation framework for tumor classification. Pattern Recognit. **93**, 293–311 (2019)
[22] Fazel, M.: Matrix rank minimization with applications. Ph.D. dissertation, Stanford University, Stanford, CA, USA (2002)
[23] Recht, B.: A simpler approach to matrix completion. J. Mach. Learn. Res. **12**, 3413–3430 (2009)
[24] Hestenes, M.: Multiplier and gradient methods. J. Optim. Theory Appl. **4**, 303–320 (1969)
[25] Yuan, X., Yang, J.: Sparse and low rank matrix decomposition via alternating direction method. Pac. J. Optim. **9**, 167–180 (2013)
[26] Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends Mach. Learn. **3**, 1–122 (2010)
[27] Gabay, G., Mercier, B.: A dual algorithm for the solution of nonlinear variational problems via finite element approximation. Comput. Math Appl. **2**, 17–40 (1976)
[28] Zhang, G., Yan, P., Zhao, H., Zhang, X.: A computer aided diagnosis system in mammography using artificial neural networks. In: IEEE International Conference on BioMedical Engineering and Informatics, vol. 2, pp. 823–826 (2008)
[29] Varela, C., Tahoces, P., Mendez, A., Souto, M., Vidal, J.: Computerized detection of breast masses in digitized mammograms. Comput. Biol. Med. **37**, 214–226 (2007)
[30] Cover, T., Hart, P.: Nearest neighbor pattern classification. IEEE Trans. Inf. Theory **13**, 21–27 (1967)
[31] Furey, T., Cristianini, N., Duffy, N., Bednarski, D., Schummer, M., Haussler, D.: Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics **16**, 906–914 (2000)
[32] Wright, J., Yang, A., Ganesh, A., Sastry, S., Ma, Y.: Robust face recognition via sparse representation. IEEE Trans. Pattern Anal. Mach. Intell. **31**, 210–227 (2009)

[33] Nasir, M., Baig, A., Khanum, A.: Brain tumor classification in MRI scans using sparse representation. In: International Conference on Image & Signal Processing, vol. 8509, pp. 629–637 (2014)
[34] Guo, Y., Wang, Y., Kong, D., Shu, X.: Automatic classification of intracardiac tumor and thrombi in echocardiography based on sparse representation. IEEE J. Biomed. Health Inform. **19**, 601–611 (2015)
[35] Zhang, L., Yang, M., Feng, X.: Sparse representation or collaborative representation: which helps face recognition? In: IEEE International Conference on Computer Vision, vol. 2011, pp. 471–478 (2012)
[36] Yang, X., Liu, F., Tian, L., Li, H., Jiang, X.Y.: Pseudo-full-space representation based classification for robust face recognition. Signal Process. Image Commun. **60**, 64–78 (2018)
[37] Lin, J.: Projected gradient methods for nonnegative matrix factorization. Neural Comput. **19**, 2756–2779 (2007)
[38] Hoyer, P.: Non-negative sparse coding. In: IEEE Workshop on Neural Networks for Signal Processing. pp. 557–565 (2004)
[39] Cai, J., Caneds, E., Shen, Z.: A singular value thresholding algorithm for matrix completion. SIAM J. Optim. **20**, 1956–1982 (2008)
[40] Strang, G.: The discrete cosine transform. SIAM Rev. **41**, 135–147 (1999)
[41] Bradley, A.P.: The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognit. **30**, 1145–1159 (1997)
[42] Kwok, J.Y.: Moderating the outputs of support vector machine classifiers. IEEE Trans. Neural Netw. **10**, 1018–1031 (1999)
[43] Vickers, A.J., Elkin, E.: Decision curve analysis: a novel method for evaluating prediction models. Med. Decis. Mak. **26**, 565–574 (2006)
[44] Yang, M., Zhang, L., Yang, J., Zhang, D.: Regularized robust coding for face recognition. IEEE Trans. Image Process. **22**, 1753–1766 (2013)
[45] Deng, W., Hu, J., Guo, J.: Extended SRC: undersampled face recognition via intraclass variant dictionary. IEEE Trans. Pattern Anal. Mach. Intell. **34**, 1864–1870 (2012)
[46] Setiawan, A.S., Wesley, J., Purnama, Y.: Mammogram classification using law's texture energy measure and neural networks. Procedia Comput. Sci. **59**, 92–97 (2015)
[47] Kutluk, S., Günsel, B.: Tissue density classification in mammographic images using local features. In: Signal Processing and Communications Applications Conference, vol. 32, pp. 1–4 (2013)
[48] Rampun, A., Scotney, B., Morrow, P., Wang, H., Winder, J.: Breast Density Classification Using Multiresolution Local Quinary Patterns in Mammograms. J. Imaging **4**, 14 (2018)
[49] Herwanto, A.M.A., Arymurthy, A.M.: Association technique based on classification for classifying microcalcification and mass in mammogram. Int. J. Comput. Sci. Issues **10**, 252–259 (2013)
[50] Golub, G.H., Loan, C.F.V.: Matrix Computations, pp. 242–243. Johns Hopkins University Press, Baltimore (1996)