An Iterated Laplacian Based Semi-Supervised Dimensionality Reduction for Classification of Breast Cancer on Ultrasound Images

Xiao Liu, Jun Shi, Shichong Zhou, Minhua Lu

Abstract—**The dimensionality reduction is an important step in ultrasound image based computer-aided diagnosis (CAD) for breast cancer. A newly proposed $l_{2,1}$ regularized correntropy algorithm for robust feature selection (CRFS) has achieved good performance for noise corrupted data. Therefore, it has the potential to reduce the dimensions of ultrasound image features. However, in clinical practice, the collection of labeled instances is usually expensive and time costing, while it is relatively easy to acquire the unlabeled or undetermined instances. Therefore, the semi-supervised learning is very suitable for clinical CAD. The iterated Laplacian regularization (Iter-LR) is a new regularization method, which has been proved to outperform the traditional graph Laplacian regularization in semi-supervised classification and ranking. In this study, to augment the classification accuracy of the breast ultrasound CAD based on texture feature, we propose an Iter-LR-based semi-supervised CRFS (Iter-LR-CRFS) algorithm, and then apply it to reduce the feature dimensions of ultrasound images for breast CAD. We compared the Iter-LR-CRFS with LR-CRFS, original supervised CRFS, and principal component analysis. The experimental results indicate that the proposed Iter-LR-CRFS significantly outperforms all other algorithms.**

I. Introduction

Breast cancer is one of the most common cancers for females worldwide. According to the statistics in 2014, it is about 29% of all new cancer cases among women in the US, which is the highest rate in all cancer types, and it is also the second leading cause of female cancer deaths [1].

Currently, ultrasound imaging has been widely used to detect breast cancer in clinic, because it is a radiation-free, effective, inexpensive and real-time imaging tool. Moreover, the ultrasound based computer-aided diagnosis (CAD) for breast cancer also attracts considerable interest, which offers more objective evaluation and improves the diagnostic accuracy and sensitivity [2][3]. Consequently, machine learning technique plays an important role in breast ultrasound CAD, among which the feature selection is one of the crucial steps [3]. However, the original extracted features usually have high dimensions with redundancy features. Therefore, dimensionality reduction (DR) methods have been widely used after feature extraction in breast ultrasound CAD.

In clinical practice, the collection of labeled instances is usually expensive, difficult or time costing, while it is relatively easy to acquire the unlabeled or undetermined instances. For example, the final diagnosis result of breast cancer usually depends on the pathology, and the ultrasound images are sometimes regarded as undetermined before pathological examination. Therefore, breast ultrasound CAD is usually a small sample size (SSS) problem with limited labeled instances by pathological examination. As a result, the semi-supervised learning (SSL), which have the ability to use unlabeled data to improve performance, have attracted much attentions for breast cancer classification now. Both semi-supervised feature selection methods and semi-supervised classification methods have been applied to breast cancer classification [4][5][6][7].

In recent years, the graph Laplacian has become an important method in manifold related machine learning, and the Laplacian regularization (LR) is widely used in SSL [8]. However, some research have indicated that when the labeled data is fixed, while the unlabeled data increases, the estimator on unlabeled points degenerates to a constant, with 'spikes' at labeled points, because the solution space is too rich, which leads to over-fitting [9][10].

To solve this problem, a regularization method using higher order Sobolev semi-norm is proposed, which uses iterated Laplacian semi-norm as this Sobolev semi-norm [10]. It can be viewed as a generalization of the thin plate spline to an unknown submanifold in high dimensions. The iterated Laplacian regularization (Iter-LR) has been proved to outperform the traditional graph Laplacian regularization in semi-supervised classification and ranking [10][12]. Iter-LR has the potential to be used in other SSL applications, such as dimensionality reduction and clustering.

The features extracted from ultrasound images for breast cancer classification usually include shape features and texture features. Therefore, the DR techniques are commonly used for these high dimensional features . Due to the SSS problem, semi-supervised DR (SSDR) technique is necessary. However, only few SSDR algorithms are applied to the features of ultrasound image for breast cancer classification. Moreover, most of the current used SSDR algorithms are based on LR, whose performance are affected by the shortcoming of LR.

Ultrasound images are affected by the speckle noise, therefore, the DR algorithms should have strong robustness

This research is supported by the Shanghai Municipal Natural Science Foundation (12ZR1410800), the Innovation Program of Shanghai Municipal Education Commission (13YZ016), and the Guangdong Key Laboratory for Biomedical Measurements and Ultrasound Imaging.

Xiao Liu is with the School of Communication and Information Engineering, Shanghai University, Shanghai, China (e-mail: liuxiao@shu.edu.cn).

Jun Shi is with the School of Communication and Information Engineering, Shanghai University, Shanghai, China (+86-21-66137256; e-mail: junshi@staff.shu.edu.cn).

Shichong Zhou is with the Department of ultrasound, Fudan University Shanghai Cancer Center; Department of Oncology, Shanghai Medical College, Fudan University, Shanghai, China (e-mail: zerglly@hotmail.com).

Minhua Lu is with the Department of Biomedical Engineering, Shenzhen University, Shenzhen, China (e-mail: luminhua@szu.edu.cn).

978-1-4244-7929-0/14/$26.00 ©2014 IEEE

against noise. Recently, a $l_{2,1}$ regularized correntropy algorithm for robust feature selection (hereafter abbreviated to CRFS) has been proposed [12]. CRFS can extract robust and sparse features, therefore, outperform some classical DR algorithms, such as principal component analysis (PCA) [12]. It is believe that CRFS has the potential to reduce features of ultrasound images. However, current CRFS is a supervised DR method, which limits its applications.

In this paper, we propose an Iter-LR-based CRFS (Iter-LR-CRFS) algorithm, which perform semi-supervised DR, and then applied it to reduce the feature dimensions of ultrasound images for classification of breast cancer.

II. METHOD

A. CRFS algorithm

The $l_{2,1}$-norm based feature selection methods often aim to solve the following constrained $l_{2,1}$-norm minimization problem [12]:

$$\min_{U} \left\| U \right\|_{2,1} \text{ s.t. } X^T U = Y \tag{1}$$

where $\| \cdot \|_{2,1}$ is an $l_{2,1}$-norm, projection matrix $U \in R^{d \times c}$, data matrix , and label matrix $Y \in R^{n \times c}$. n, d and c are the number of training samples, feature dimension, and classes, respectively.

The half-quadratic (HQ) minimization is an effective optimization technique with successful applications in computer vision [12][13]. Considering the HQ analysis for $l_{2,1}$-norm, the following general robust learning problem is considered [12]:

$$\min_{U} \sum_{i=1}^{d} \varphi_o \left(\left\| (AU + B)^i \right\|_2 \right) + \gamma \sum_{i=1}^{d} \varphi_R \left(\left\| u^i \right\|_2 \right) \tag{2}$$

Correntropy is proposed in information theoretic learning to process non-Gaussian noise [14], and has been successfully used in computer vision [12][14]. By applying correntropy and $l_{2,1}$-norm regularization in (2), the objective of CRFS is obtained [12]:

$$\min_{U} \{ 1 - \sum_{k=1}^{n} \exp \left(- \left\| (X^T U - Y)^k \right\|_2^2 / \delta^2 \right) + \left\| U \right\|_{2,1} \} \tag{3}$$

where σ is the kernel size that controls all properties of correntropy. Here, correntropy is used to remove outliers and $l_{2,1}$-norm regularization to select robust and informative features. HQ method is then successfully adopted to solve the optimization of (3) [12].

The CRFS method can extract informative and discriminative features, and remove the irrelevent and redundant features by minimizing objective function. Therefore the CRFS method is robust to the outlier and noise.

B. Iterated Laplacian regularization

In LR-based SSL, the typical form of the optimization problem is

$$\min_{f} \sum_{x_i \in X_L} (f(x_i) - y_i)^2 + \mu f^T L f \tag{4}$$

where $f^T L f$ is Laplacian regularization item, and it is defined as following:

$$f^T L f = \frac{1}{2} \sum_{x_i, x_j \in X} w_{ij} (f(x_i) - f(x_j)) \tag{5}$$

where X is the dataset including both labeled and unlabeled samples, W_{ij} is a similarity weight between sample x_i and x_j.

The Iter-LR method proposed by Zhou et al. achieves better performance than LR method [10]. Iter-LR equals to a high order Sobolev semi-norm, and its theory is introduced as following [10]:

Define the iterated Laplacian semi-norm as

$$I_m^d (f) = \int_{\Omega} f(x) \Delta^m f(x) dx \tag{6}$$

where f is the projection of a continuous function f(x) on the sampel set X , m is the order of semi-norm, and Ω is a compact Riemannian submanifold. Then, the empirical version of (6) is given by

$$I_{m,n}^d (f) = f^T L^m f \tag{7}$$

where $I_{m,n}^d (f)$ is a semi-norm without further conditions, L is graph Laplacian, and n means that L is built on total n data points. When Ω has a smooth boundary , the limit of $I_{m,n}^d (f)$ is the same given proper boundary conditions. Increasing m restricts solution space to be a smoother space, and from kernel point of view, increasing m corresponds to a better density adaptive kernel.

Here, the iterated Laplacian semi-norm $I_{m,n}^d (f)$ is used as the Sobolev semi-norm, which corresponds to the empirical iterated Laplacian regularizer $f^T L^m f$ given finite data, and also has the advantage of being coordinate free.

Iter-LR based SSL requires only a trivial modification of the optimization problem LR-SSL. The optimization problem of Iter-LR-based SSL is then given by

$$\min_{f} \sum_{x_i \in X_L} (f(x_i) - y_i)^2 + \mu I_{m,n}^d (f) \tag{8}$$

C. Iterated Laplacian based semi-supervised CRFS

The current CRFS is a supervised algorithm. Since SSL can improve the learning task with the help of unlabeled samples, and Iter-LR performs well in SSL, we propose the Iter-LR-CRFS.

The objective of LR-based CRFS (LR-CRFS) is restricted by LR in (5) as a penalty term, given by

$$\begin{aligned} \min_{U} \{ & 1 - \sum_{k=1}^{n} \exp \left(- \left\| (X^T U - Y)^k \right\|_2^2 / 2 \right) + \left\| U \right\|_{2,1} \\ & + \lambda \sum_{x_i, x_j \in X} w_{ij} (f(x_i) - f(x_j)) \} \end{aligned} \tag{9}$$

The iterated Laplacia matrix means the power of Laplacian matrix, which is simply improved on Laplacian regularization [10]. Therefore, the objective of Iter-LR-CRFS

is penalized by the iterated Laplacian regularization item as following:

$$\min_{U} \{1 - \sum_{k=1}^{n} \exp\left(-\left\|(X^T U - Y)^k\right\|_2^2/2\right) + \|U\|_{2,1} + \lambda I_m^d\} \quad (10)$$

The empirical version $I_{m,n}^d(f)$ of the iterated Laplacian is used instead of $I_m^d(f)$, because $I_{m,n}^d(f)$ is converged to $I_m^d(f)$ when f is smooth function on probably.

We use an alternate minimization way to solve the objective function (10). According to HQ optimization method, the auxiliary variables p and q of correntropy are inducted as following:

$$\begin{aligned} p_k^t &= \exp\left(-\left\|(X^T U - Y)^k\right\|_2^2/\sigma^2\right) \\ q_i^t &= 1/(2\left\|u^i\right\|_2) \\ U^t &= \arg\min_{U} (\mathrm{Tr}(X^T U - Y)^T P(X^T U - Y) + \\ &\quad \lambda_1 \mathrm{Tr}(U^T Q U) + \lambda_2 I_{m,n}^d(f)) \end{aligned} \quad (11)$$

where P=diag(p) and Q=diag(q).

The analytic solution of (11) is given by

$$U^* = (XPX^T + \lambda_1 Q + \lambda_2 X L^m X)^{-1} XP Y \quad (12)$$

The above-mentioned algorithm is the proposed Iter-LR-CRFS.

III. Experiment

A. Data

To evaluate the performance of proposed Iter-LR-CRFS algorithm, we applied it to breast ultrasound images for dimensionality reduction.

A total of 200 pathology-proved breast ultrasound images (including 100 benign masses and 100 malignant tumors) were used to evaluate the performance of the proposed method. These images were randomly selected from the Cancer Hospital of Fudan University by one of the authors. These images were acquired from several ultrasound imaging devices by different manufacturers. The human subject ethical approval was obtained from the relevant committee in the Cancer Hospital of Fudan University before carrying out the experiment. Each subject provided a written consent prior to the experiment. To select the ROI, the position of mass was roughly located at the center of the original breast ultrasound image, and then the image was cropped to the size of 128×128.

The Shearlet-based texture features were extracted from ultrasound images, which has proved the effectiveness in our previous work [15]. The dimensions of the Shearlet-based texture features is 148.

B. Experimental Setup

We compared the proposed Iter-LR-CRFS with LR-CRFS, original CRFS and the original Shearlet-based texture features. Moreover, since PCA is the most commonly used DR algorithm, it is also performed for comparison here. The linear SVM in LIBSVM software was adopted as the classifier in this study [15].

The 4-fold cross-validation strategy was performed on the 200 samples. Two fold samples were used as unlabeled data, and the remaining two fold samples were the training data and testing data, respectively. To keep the balance between benign and malignant image for the classifier, the ratios of benign images to malignant images was 1:1 in each fold.

The classification accuracy, sensitivity and specificity were selected as the evaluation indices, which were commonly used in medical image and signal classification. A paired-samples *t*-test was used to statistically evaluate the performances between the proposed Iter-LR-CRFS and other DR algorithms. The results were declared statistically significant when associated with *p*-value that is less than 0.05.

All the algorithms were implemented using MATLAB 2009b (MathWorks, Massachusetts, USA) and performed on a Dell computer (2.66GHz/2G RAM).

IV. Experimental Results

Table 1 show the classification results of different feature extraction algorithms. 100 unlabeled samples were used in semi-supervised Iter-LR-CRFS and LR-CRFS. It can be found that the proposed Iter-LR-CRFS achieves best performance on classification accuracy and sensitivity, which are 89.0±3.6% and 91.0±5.2%, respectively. While the LR-CRFS get best result on specificity (91.0±6.6%). Iter-LR-CRFS significantly outperforms other compared algorithms on classification accuracy and sensitivity with all the *p*-values less than 0.05. Specifically, Iter-LR-CRFS provides at least 1.5% and 7.0% improvements on accuracy and sensitivity compared with LR-CRFS, and 3.5% and 9% improvements compared with original CRFS, correspondingly. The specificity of Iter-LR-CRFS is lower than LR-CRFS and CRFS, because the higher classification accuracy of Iter-LR-CRFS is to obtained at the cost of specificity. It also found that both the semi-supervised based CRFS (Iter-LR-CRFS and LR-CRFS) outperform supervised CRFS algorithms and unsupervised PCA algorithm in this work, because more unlabeled data improve the performance in SSL.

Figure 1 shows the classification accuracies of semi-supervised CRFS algorithms with respect to different number of unlabeled samples used for helping training. We added the unlabeled samples from 20 to 100 with an interval of 20. It can be found that the classification accuracies of both Iter-LR-CRFS and LR-CRFS steadily improve with the increase of unlabeled samples. Moreover, Iter-LR-CRFS always outperforms LR-CRFS with different unlabeled samples, because the solution of Iter-LR is more stable than that of LR in finite unlabeled data case.

It is worth noting that the reduced features by CRFS and semi-supervised CRFS are two dimensions, which is much less than the original 148 texture features. The original Shearlet-based texture feature has been proved the discriminative representativeness and effectiveness in breast ultrasound CAD in our previous work [15]. The original texture features are relatively compact without much redundant information, because the features extracted by PCA reduce the classification performance on the contrary. However, the reduced features by CRFS achieve 1% improvement of classification accuracy compared with original texture features, but with much lower feature dimension. One of the reasons is that CRFS has stronger robustness against noise than PCA. Therefore, CRFS is suitable DR method for of ultrasound features, which are corrupted by speckle noise in ultrasound image. Moreover, the less features can benefit the computational efficiency of classifier, especially for large-scale data classification. While with the help of unlabeled samples, semi-supervised CRFS further improve the performance.

TABLE I. CLASSIFICATION RESULTS OF DIFFERENT FEATURE SELECTION METHODS (UNIT: %)

	ACC	SEN	SPE
148 Features	84.5±5.7	80.0±8.5	89.0±4.4
PCA	83.5±5.2	78.0±3.5	89.0±7.7
CRFS	85.5±3.9	82.0±4.5	89.0±5.9
LR-CRFS	87.5±3.6	84.0±2.8	**91.0±6.6**
Iter-LR-CRFS	**89.0±3.6**	**91.0±5.2**	87.0±9.1

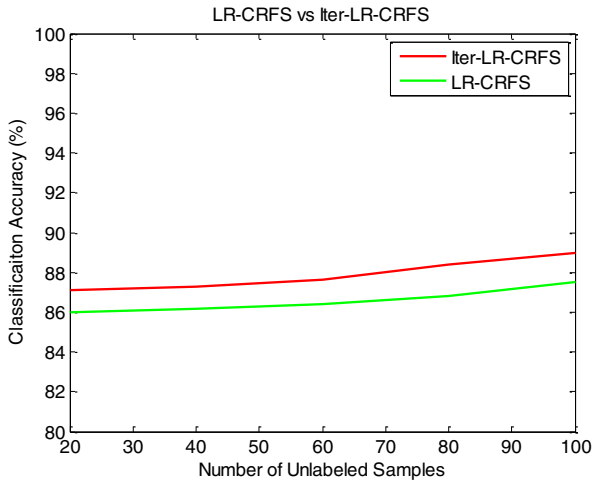

Figure 1. Classification accuracies of Iter-LR-CRFS and LR-CRFS with respect to the different number of unlabeled samples

V. CONCLUSION

In this paper, we proposed a iterated Laplacian regularization based semi-supervised CRFS algorithm to reduce the feature dimension of ultrasound images. The experimental results indicate that our proposed Iter-LR-CRFS algorithm can significantly improve the performance of extracting discriminative features from ultrasound images for breast cancer classification, which suggests that it has the potential to be used in ultrasound based CAD. In the future work, the multi-view Iter-LR-CRFS algorithm will be studied to reduce the dimensions of shape features and texture features simultaneously for further improve the performance of breast ultrasound CAD.

REFERENCES

[1] R. Siegel, J. Ma, Z. Zou, A. Jemal. "Cancer statistics," *CA Cancer J. Clin*., vol. 64, pp. 9-29, 2014.

[2] C. M. Sehgal, S. P. Weinstein, P. H. Arger, F. E. Conant. "A review of breast ultrasound," *J. Mammary Gland Biol.*, vol. 11, pp. 113-123, 2006.

[3] H. D. Cheng, J. Shan, W. Ju, et al. "Automated breast cancer detection and classification using ultrasound images: a survey," *Pattern Recogn*., vol. 43, pp. 299-317, 2010.

[4] J. Li, Y. Yu, Z. Yang, L. Tang. "Breast tissue image classification based on semi-supervised locality discriminant projection with kernels," *J. Med. Syst*., vol. 36, pp. 2779-2786, 2012.

[5] J. Kim, H. Shin. "Breast cancer survivability prediction using labeled, unlabeled, and pseudo-labeled patient data," *J. Am. Med. Inform. Assoc*., vol. 20, pp. 613-618, 2013.

[6] Y. Nam, H. Shin. "A hybrid cancer prognosis system based on semi-supervised learning and decision trees," *Neural Inform. Process*., vol. 8227, pp. 640-648, 2013.

[7] X. Liu, J. Liu, Z. Feng, et al. "Mass classification in mammogram with semi-supervised relief based feature selection," In *15th Int. Conf. Graph. Image Process.*, 2013.

[8] M. Belkin, P. Niyogi, S. Sindhwani. "Manifold regularization: a geometric framework for learning from labeled and unlabeled examples," *J. Mach. Learn. Res.*, vol. 7, pp. 2399-2434, 2006.

[9] B. Nadler, N. Srebro, X. Zhou. "Statistical analysis of semi-supervised learning: the limit of infinite unlabelled data," In *Adv. Neural Inf. Process. Syst.*, vol. 22, pp. 1330-1338, 2009.

[10] X. Zhou, M. Belkin. "Semi-supervised learning by higher order regularization," In *14th Int. Conf. Artif. Intel. Stat.*, pp. 892-900, 2011.

[11] X. Zhou, M. Belkin, N. Srebro. "An iterated graph Laplacian approach for ranking on manifolds," In *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 877-885, 2011.

[12] R. He, T. Tan, L. Wang, W. Zheng, "$l_{2,1}$ regularized correntropy for robust feature selection," In *IEEE Conf. Comput. Vis. Pattern Recog*., pp. 2504-2511, 2012.

[13] R. He, W. Zheng, T. Tan, Z. Sun, "Half-quadratic based iterative minimization for robust sparse representation," *IEEE Trans. Pattern Anal. Mach. Intel.*, vol. 36, pp. 261-275, 2013.

[14] W. Liu, P. P. Pkharel, J. C. Principe, "Correntropy: properties and applications in non-gaussian signal processing," *IEEE Trans. Signal Process.*, vol. 55, pp.5286-5298, 2007.

[15] S. Zhou, J. Shi, J. Zhu, et al. "Shearlet-based texture feature extraction for classification of breast tumor in ultrasound image," *Biomed. Signal Process. Control*, vol. 8, pp. 688-696, 2013.

[16] C. C. Chang, C. J. Lin, "LIBSVM: a library for support vector machines," 2001.