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

Computerized breast cancer analysis system using three stage semi-supervised learning method

Wenqing Sun [a], Tzu-Liang (Bill) Tseng [b], Jianying Zhang [c,d], Wei Qian [a,d,*]

[a] Department of Electrical and Computer Engineering, University of Texas at El Paso, 500 West University Avenue, El Paso, TX 79968, USA

[b] Department of Industrial, Manufacturing and Systems Engineering, University of Texas at El Paso, 500 West University Avenue, El Paso, TX 79968, USA

[c] Department of Biological Sciences, University of Texas at El Paso, 500 West University Avenue, El Paso, TX 79968, USA

[d] Sino-Dutch Biomedical and Information Engineering School, Northeastern University, No.11, Lane 3, Wenhua Road, Heping District, Shenyang, Liaoning 110819, China

ARTICLE INFO

Article history:
Received 12 October 2015
Received in revised form
3 June 2016
Accepted 4 July 2016

Keywords:
Semi-supervised learning
Computer aided detection
Mass detection
Unlabeled data

ABSTRACT

Background and Objective: A large number of labeled medical image data is usually a requirement to train a well-performed computer-aided detection (CAD) system. But the process of data labeling is time consuming, and potential ethical and logistical problems may also present complications. As a result, incorporating unlabeled data into CAD system can be a feasible way to combat these obstacles.

Methods: In this study we developed a three stage semi-supervised learning (SSL) scheme that combines a small amount of labeled data and larger amount of unlabeled data. The scheme was modified on our existing CAD system using the following three stages: data weighing, feature selection, and newly proposed dividing co-training data labeling algorithm. Global density asymmetry features were incorporated to the feature pool to reduce the false positive rate. Area under the curve (AUC) and accuracy were computed using 10 fold cross validation method to evaluate the performance of our CAD system. The image dataset includes mammograms from 400 women who underwent routine screening examinations, and each pair contains either two cranio-caudal (CC) or two mediolateral-oblique (MLO) view mammograms from the right and the left breasts. From these mammograms 512 regions were extracted and used in this study, and among them 90 regions were treated as labeled while the rest were treated as unlabeled.

Results: Using our proposed scheme, the highest AUC observed in our research was 0.841, which included the 90 labeled data and all the unlabeled data. It was 7.4% higher than using labeled data only. With the increasing amount of labeled data, AUC difference between using mixed data and using labeled data only reached its peak when the amount of labeled data was around 60.

Conclusions: This study demonstrated that our proposed three stage semi-supervised learning can improve the CAD performance by incorporating unlabeled data. Using unlabeled data is promising in computerized cancer research and may have a significant impact for future CAD system applications.

[*] *Corresponding author.* Department of Electrical and Computer Engineering, Medical Imaging Informatics Laboratory, College of Engineering, University of Texas at El Paso, 500 West University Avenue, El Paso, TX 79968, USA. Fax: +1 (915) 747 7871.
E-mail address: wqian@utep.edu (W. Qian).

http://dx.doi.org/10.1016/j.cmpb.2016.07.017
0169-2607/© 2016 Elsevier Ireland Ltd. All rights reserved.

1. Introduction

Breast cancer refers to the erratic growth of cells that originate in breast tissue and is one of the most common types of cancer [1]. Scientific evidences have shown that early cancer detection is important to enhance the survival rate of the patients through more effective patient management and treatment [2,3]. Mammography has been the most cost-effective and widely used imaging modality for breast cancer screening during last few several decades [4]. To help the radiologists diagnose early breast cancer, many methodologies have been implemented to improve the performance of computer aided detection (CAD) systems.

However, most of the existing CAD systems are implemented via supervised learning which is based on a large volume of labeled breast cancer image data. These data come with truth files (disease statuses marked by the radiologists such as “malignant” or “benign”) and to obtain them requires radiologists to read and mark the mammograms. The malignant masses are marked and labeled as malignant areas while the rest of the areas are considered as benign or normal. In CAD systems, features are usually extracted from region of interest (ROI), so every ROI is condensed into a vector of numerical numbers and the labels of corresponding areas will be paired with these features. Unfortunately, acquiring a large amount of labeled data is usually time consuming [5], and making diagnosis on such large number of cases is a heavy burden for radiologists. In order to label all the samples, several radiologists are required to evaluate all the data individually, compare their evaluations with one another, and conduct a discussion of the resulting analysis to reach a final conclusion. Moreover, confidentiality agreements are required to obtain the data, and typically neither the doctors nor the patients are willing to reveal this information. On the other hand, abundance of unlabeled data (no truth file related or “truth partially known”) is likely to be more accessible in most research contexts and will significantly reduce the problem of confidential sensitivity [6].

One feasible way to solve this problem is to combine a small amount of labeled data with a large amount of unlabeled data to build the classifier together. In the area of machine learning, this technique is called semi-supervised learning (SSL). It exploits labeled data and unlabeled data at the same time without any human intervention. The SSL algorithm allows us to label some of the unlabeled data based on the information provided by the initial labeled dataset, and the newly labeled data (also called pseudo-labeled data [7]) can be used to improve the CAD performance. One popular and promising SSL method is called co-training, which is proposed by Blum and Mitchell [8]. It trains two classifiers by letting each classifier label the unlabeled data for the other one, and it makes decisions based on the agreement of these two classifiers. Although the co-training method has been successfully used in many fields [9–11], the requirement of two sufficient and redundant feature subsets can be hardly met in CAD systems. In the breast cancer CAD area, some research groups have tried different SSL methods to utilize the unlabeled data [12], extended the co-training algorithm by incorporating the random forest ensemble method to determine the most confident examples to label [6], and made a comparison of the transductive dimension reduction method and semi-supervised LapSVM manifold regularization method with an ultrasound image database containing 1126 lesions. In our group, we have conducted a co-training based SSL breast cancer research based on labeled and unlabeled ROIs [13].

The need to avoid false positives cannot be over stressed for CAD systems since it might bring psychological harms, unnecessary imaging tests and biopsies in women without cancer, and inconvenience [14,15]. Moreover, different from traditional supervised machine learning techniques, SSL methods have to overcome several other obstacles including the conflict of accuracy and diversity [16] as well as noisy data in labeled dataset [17]. Many efforts have been made to solve these problems. For example, to reduce the false positive rate, a number of computerized schemes have been reported and tested to segment fibroglandular tissue and compute mammographic density [18–20], and Wang et al. [4] found that the incorporation of global mammographic density measurements into a CAD system can help accomplish this goal. To create the diversity, Jiang et al. [16] developed a method by manipulating the training set, and based on this theory they proposed the inter-training algorithm. To eliminate the noisy data, instant selection algorithms have been discussed and compared [17,18]. However, to the best of our knowledge, no mammogram based SSL scheme has been reported, and the possibility of redesigning and modifying the traditional supervised learning based CAD systems to semi-supervised learning CAD systems have never been investigated and tested.

The aim of this study is to investigate and test that whether redesigning CAD system and incorporating the modified SSL algorithm can efficiently use unlabeled data and thus improve the performance of detection result. For this purpose, a novel three stage semi-supervised learning approach was designed and tested based on our existing CAD system [19–21]. This system has been widely tested in the clinical practice to assist radiologists in reading and interpreting mammograms to date [22,23]. Our SSL method integrated data weighing, feature selection and a newly proposed dividing co-training data labeling method. The global density asymmetry features were integrated into the CAD system, and each module of the system was modified and redesigned to adapt to the proposed method. Dr. Wei Qian organized this study, and the data we used for this study were collected by Dr. Wei Qian, Dr. Tzu-Liang (Bill) Tseng and Dr. Jianying Zhang. The method was developed by Wenqing Sun and Dr. Wei Qian together.

2. Materials and methods

2.1. Data

From an established in-house full-field digital mammography (FFDM) image database, 400 cases were investigated and tested which included 200 pairs of cranio-caudal (CC) view and 200 pairs of mediolateral-oblique (MLO) view breast images, and each pair contains two mammograms acquired from both right and left breasts. The women’s ages range from 32 to 68 years, with a mean age of 47.7 years and median age of 43 years. Fig. 1 shows that the cases in positive and negative