Computers in Biology and Medicine 83 (2017) 157–165

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

A novel computer-aided diagnosis system for breast MRI based on feature selection and ensemble learning

Wei Lu, Zhe Li, Jinghui Chu*

School of Electronic Information Engineering, Tianjin University, Tianjin 300072, PR China

ARTICLE INFO

Keywords:
Computer-aided diagnosis
Breast cancer
MRI
Feature selection
Classification
Ensemble learning

ABSTRACT

Breast cancer is a common cancer among women. With the development of modern medical science and information technology, medical imaging techniques have an increasingly important role in the early detection and diagnosis of breast cancer. In this paper, we propose an automated computer-aided diagnosis (CADx) framework for magnetic resonance imaging (MRI). The scheme consists of an ensemble of several machine learning-based techniques, including ensemble under-sampling (EUS) for imbalanced data processing, the Relief algorithm for feature selection, the subspace method for providing data diversity, and Adaboost for improving the performance of base classifiers. We extracted morphological, various texture, and Gabor features. To clarify the feature subsets' physical meaning, subspaces are built by combining morphological features with each kind of texture or Gabor feature. We tested our proposal using a manually segmented Region of Interest (ROI) data set, which contains 438 images of malignant tumors and 1898 images of normal tissues or benign tumors. Our proposal achieves an area under the ROC curve (AUC) value of 0.9617, which outperforms most other state-of-the-art breast MRI CADx systems. Compared with other methods, our proposal significantly reduces the false-positive classification rate.

1. Introduction

Breast cancer is a disease that is caused by malignant cells in the breast tissues [1]. According to American National Cancer Institute, the incidence rate of female breast cancer was 124.8 per 100,000 women per year from 2008 to 2012. In 2015, it is estimated that there were 231,840 new breast cancer cases and about 40,290 people died from this disease. In the U.S., female breast cancer is responsible for 14% of all new cancer cases [2], indicating that female breast cancer is the most common cancer. The mortality rate from breast cancer is also the highest among women [3]. Breast cancer pathogenesis remains unknown, and there is no effective way to prevent this disease [4]. However, early detection and diagnosis of breast cancer can help to significantly reduce the mortality rate. Thus, much research has been performed on the early detection of malignant masses. Modern imaging techniques including ultrasound, mammography, computed tomography (CT), positron emission tomography (PET), and magnetic resonance imaging (MRI) have been widely used for the early detection and diagnosis of breast cancer [5]. Among these techniques, MRI is well-known for its superiority in prognosis, diagnostic accuracy, staging, and preoperative planning [67]. It was also shown that MRI has better sensitivity than mammography and MRI diagnostic results are only minimally influenced by breast density [8]. Therefore, MRI is considered to be an important tool in breast cancer clinical diagnosis [5].

The goal of Computer-Aided Detection and Diagnosis (CAD) is to achieve a high diagnostic sensitivity for breast cancer and to maintain a low the false positive classification (FPC) rate [5]. In this process, Computer-Aided Diagnosis (CADx) is regarded as a key technique to reduce the FPC rate [9].

An important task of CADx is to make an accurate mass classification and decide whether a region of interest (ROI) is malignant. Enough high-quality features that characterize malignant masses are needed for training the classification model and for class prediction [10]. Thus, many features, such as morphological features, texture features, and frequencial features, have been extracted and used widely in many studies [11,12,13,14]. However, some extracted features may be redundant and irrelevant to the classification task. In addition, too many input features may increase the computational complexity and cause the curse of dimensionality, thereby significantly diminishing the diagnostic accuracy. Thus, feature selection plays a crucial role in improving CADx system performance. Genomic algorithm (GA) [15] and support vector machine-based recursive feature elimination (SVM-RFE) [9] have been adopted in the feature selection process for CADx systems.

* Corresponding author.
E-mail addresses: luwei@tju.edu.cn (W. Lu), tywzlizhe29121@126.com (Z. Li), cjh@tju.edu.cn (J. Chu).

http://dx.doi.org/10.1016/j.compbiomed.2017.03.002
Received 30 September 2016; Received in revised form 25 February 2017; Accepted 1 March 2017

In addition to acquiring representative features of breast masses, it is also important to build and train a robust classifier. According to our knowledge, most popular classifiers are designed under the assumption that the data set used for training is balanced, which means that the number of samples in the majority class is similar to that in the minority. However, this prerequisite is difficult to achieve in CADx. There are usually much fewer images with malignant masses than those without masses or with only benign masses [16]. This may reduce the diagnostic sensitivity, and malignant masses are likely to be wrongly classified as being normal. Consequently, it is necessary to alleviate the influence caused by data imbalance. Ensemble learning algorithms such as Ensemble of Under-sampled SVM (EUS-SVM) [17] and RUSBoost [18] have been shown to perform well in breast cancer CAD [16] and other medical classification systems [19].

In this paper, we propose a MRI CADx system focusing on diagnosis of malignant breast masses, based on feature selection and ensemble learning. First, morphological, texture, and Gabor features were extracted to characterize breast cancer masses. We then used the Relief algorithm [20] to find the optimal feature subset for the classifier training. These features were then fed to a novel ensemble learning framework based on the combination of EUS and the subspace technique. The experimental results indicate that our proposal outperforms the other state-of-the-art methods in diagnostic sensitivity, but the FPC rate increases slightly.

The main contributions of this paper can be summarized as follows:

1. The dimensionality of the features we use is larger than most state-of-the-art methods. Various features including morphological, Gabor, and several types of texture features were extracted to comprehensively characterize breast masses.
2. We selected the optimal feature subset from the original feature set using Relief, based on their type, which helps reduce the redundant and irrelevant features and takes the physical meaning of features into consideration.
3. We propose a novel ensemble learning framework based on the combination of EUS, subspace, and Adaboost, which helps to alleviate the data imbalance problem and improves the overall classification accuracy of the CADx system.

The remainder of this paper is organized as follows: Section 2 describes our methodology in detail; Section 3 introduces our data set, the experimental setup, and our performance evaluation metrics; Section 4 presents the experimental results and demonstrates the effectiveness of each individual component of our proposal; Section 5 discusses the reason for our system's superiority; and Section 6 includes concluding remarks (Fig. 1).

2. Methodology

In this section, we discuss the methodology used in our proposal. The structure of our CADx system is shown in Fig. 2. Because automated ROI segmentation may cause some errors [21], here, we segmented the ROIs from breast MRI images manually with the aid of physicians' marks. In subsection A, we present the features that we extracted for characterization of breast cancer images. In subsection B, we introduce our method for feature selection and the process of subspace construction. In subsection C, we briefly introduce our ensemble learning framework for imbalanced data processing. In subsection D, we introduce our base classification algorithm. Finally, in subsection E, we present the pseudo-code for our proposal.

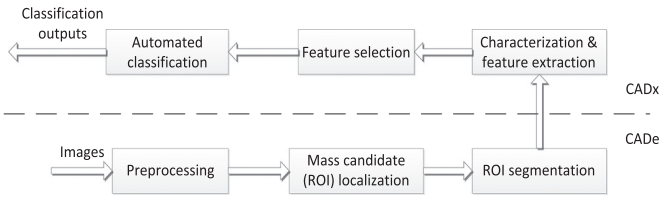

Fig. 1. This is a diagram of CAD scheme. The lower part below the dashed line shows the components of CADe, and the upper part above the dashed line shows the components of CADx.

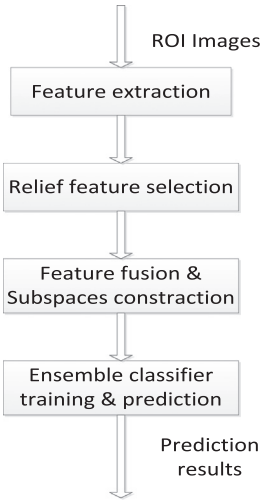

Fig. 2. The general structure of our CADx system.

2.1. Feature extraction

In the pattern recognition and image processing fields, the aim of feature extraction is to acquire numeric variables that can reflect the intrinsic image properties [22]. Malignant breast masses usually have a spiculated, rough, and blurry boundary, while benign masses usually have a round, smooth, and well-defined boundary [23]. These morphological features are crucial to the classification. However, not all morphology information is suitable for describing the ROI images. Texture and frequencial features that focus more on the tissue composition are also useful and effective for characterizing breast masses. In this study, we extract various types of features, including morphological features, gray level co-occurrence matrix (GLCM, also known as Haralick features [24]), gray level difference matrix (GLDM), gray level run length matrix (GLRLM), gradient-gray level co-occurrence matrix (GGLCM) and Gabor features, for the additional classification task. Among these features, GLCM, GLDM, GLRLM, and GGLCM features belong to the texture statistical feature family, and Gabor features belong to the frequency domain feature family. A brief introduction to these features is provided below.

2.1.1. Morphological features

Morphological features are variables that describe the shape, edge, and geometric properties of masses [25]. In our proposal, circularity, mean and standard deviation of the radial length, eccentricity, entropy of the intensity distribution, mean and standard deviation of the intensity, area, mean and standard deviation of the fractal dimension index, inertial momentum, anisotropy, entropy of the contour gradient, smoothness, skewness and kurtosis were extracted as morphological features. The dimensionality of our implementation in this part is 16.

2.1.2. Haralick features

Haralick features [24] are statistical texture features based on GLCM. Unlike the gray level histogram which can only reflect the gray level and its spatial distribution, GLCM can display information about the relative position of pixel pairs with different distances and angular relationships in the image [26]. For each breast MRI image, we extracted the following 11 features from its GLCM; namely angular second moment, contrast, correlation, difference moment, homogene-