COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE 134 (2016) 259–265

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

An immune-inspired semi-supervised algorithm for breast cancer diagnosis ☆

Lingxi Peng [a], *Wenbin Chen* [b], *Wubai Zhou* [c], *Fufang Li* [b], *Jin Yang* [d], *Jiandong Zhang* [d,*]

[a] *School of Mechanical and Electrical Engineering, Guangzhou University, Guangzhou, 510006, China*
[b] *School of Computer Science and Education, Guangzhou University, Guangzhou, 510006, China*
[c] *School of Computing & Information Sciences, Florida International University, Miami, FL 33199, USA*
[d] *Department of Computer Science, Leshan Normal Univ., Leshan 614000, China*

ARTICLE INFO

Article history:
Received 23 November 2015
Received in revised form
28 May 2016
Accepted 6 July 2016

Keywords:
Breast cancer diagnosis
Artificial immune
Machine learning

ABSTRACT

Breast cancer is the most frequently and world widely diagnosed life-threatening cancer, which is the leading cause of cancer death among women. Early accurate diagnosis can be a big plus in treating breast cancer. Researchers have approached this problem using various data mining and machine learning techniques such as support vector machine, artificial neural network, etc. The computer immunology is also an intelligent method inspired by biological immune system, which has been successfully applied in pattern recognition, combination optimization, machine learning, etc. However, most of these diagnosis methods belong to a supervised diagnosis method. It is very expensive to obtain labeled data in biology and medicine. In this paper, we seamlessly integrate the state-of-the-art research on life science with artificial intelligence, and propose a semi-supervised learning algorithm to reduce the need for labeled data. We use two well-known benchmark breast cancer datasets in our study, which are acquired from the UCI machine learning repository. Extensive experiments are conducted and evaluated on those two datasets. Our experimental results demonstrate the effectiveness and efficiency of our proposed algorithm, which proves that our algorithm is a promising automatic diagnosis method for breast cancer.

1. Introduction

Breast cancer develops from breast tissue and usually begins with the formation of a small, confined tumor (lump) or as calcium deposits (micro-calcifications), and then spreads through channels within the breast to the lymph nodes or through the blood stream to other organs. Moreover, breast cancer is an uncontrolled growth of breast cells and usually can only be found after symptoms emerge, but many women in early breast cancer have no symptoms. Nearly one in eight U.S. women have breast cancer some time during their lives [1]. Sometimes the cause of breast cancer is unknown to doctors. Early detection can be a big plus in treating breast cancer. The earlier breast cancer is detected, the higher probability it may be cured.

☆ This work was supported by the National Natural Science Foundation of China under Grant No. 61472092, No. 11271097 and No. 61100150, Creation Team Construction Project of Guangdong Province University under Grant No. 2015KCXTD014, Scientific Research Fund of Sichuan Provincial Education Department No.13TD0014, and Scientific Research Fund of Leshan Normal University under Grant No. Z1412.

* *Corresponding author.* Department of Computer Science, Leshan Normal Univ., Leshan 614000, China. Fax: +86 833 2276382.
E-mail address: zjd2003@163.com; flyingday@139.com (J. Zhang).

http://dx.doi.org/10.1016/j.cmpb.2016.07.020
0169-2607/

It has been proven that the introducing machine learning techniques into medical breast cancer analysis process gains a lot benefits. It increases diagnostic accuracy, cuts down costs and reduces human resources. In related works, many researchers try to gain more accurate breast cancer diagnosis through the use of various methods. Abbass [2] presented an evolutionary artificial neural network approach based on the pareto-differential evolution (PDE) algorithm augmented with local search for the prediction of breast cancer. Zheng et al. [3] developed a hybrid of K-means and support vector machine algorithms. The K-means algorithm is utilized to recognize the hidden patterns of the benign and malignant tumors separately. Then, a support vector machine is used to obtain the new classifier to differentiate the incoming tumors. Their proposed methodology improves the accuracy to 97.38%, when it is tested on the Wisconsin Diagnostic Breast Cancer (WDBC) data set. Li et al. [4] proposed a novel supervised dimensionality reduction method named as quasiconformal kernel common locality discriminant analysis (QKCLDA), which obtained 96.98% accuracy. Quinlan [5] presented C4.5 decision tree method and reached 94.74% classification accuracy using 10-fold cross-validation with WDBC data set. Guo et al. [6] developed a new feature extraction measure method on breast cancer diagnosis, which obtained the accuracy of 97.5%. Pena-Rayes and Sipper [7] combined two methodologies-fuzzy systems and evolutionary algorithms (FG) so as to automatically produce diagnostic systems acquiring the accuracy of 97.8% on the Breast Cancer Wisconsin (Original) (BCWO) dataset. Abonyi and Szeifert [8] also applied the supervised fuzzy clustering technique on the same one and obtained 95.57% accuracy. Bhardwaj and Tiwari [9] proposed a genetically optimized neural network method for breast cancer diagnosis. Results have showed that their approach works well on the breast cancer database. In Karabatak's research [10], a new NB (weighted Naïve Bayesian) classifier was proposed and its application on breast cancer detection was presented. Sheikhpour et al. [11] hybridized the particle swarm optimization method with the non-parametric kernel density estimation based classifier method for the diagnosis of breast cancer. Purwar and Singh presented a hybrid prediction model with missing value imputation (HPM-MI) using simple K-means clustering on BCWO dataset [12].

These techniques mentioned above belong to the supervised learning methods. Supervised learning techniques only use the labeled sample sets and unsupervised learning methods only use unlabeled sample sets. However, in practice, only a small amount of labeled data is available since it is very expensive to label the data. In biology and medicine, unlabeled data is very easy to access. Thus, the semi-supervised learning techniques have a promising prospect in medical diagnosis. In this paper, we adapt semi-supervised learning techniques to utilize the large number of unlabeled data.

Biological immune system has many desirable properties. It is versatile, efficient and robust to errors. Furthermore, it can learn to recognize the antigens of specific pathogens and remember them for the future. In the last twenty years, there has been a great deal of interest in exploiting the known properties of the biological immune theory as metaphorical inspiration for solving computational problem. Exciting results have been obtained from the research area in machine learning [13], combination optimization [14], pattern recognition, etc. [15,16]. It has been also successfully applied in the field of disease diagnosis [17,18], and delivers promising diagnostic accuracies. Among these research, artificial immune recognition system (AIRS) has shown significant success on a broad range of classification problems, which has been successfully applying to the breast cancer problem. Saybani et al. incorporated support vector machine, fuzzy logic, and real tournament selection mechanism into AIRS. They obtained 100% accuracy on BCWO dataset [19]. Kung Jeng Wang et al. combined the Synthetic Minority Over-Sampling Technique (SMOTE) with AIRS on the WDBC and BCWO datasets. They improved the accuracy to 96.91% and 96.52%, respectively [20]. In Polata et al.'s research, resource allocation mechanism of AIRS was changed with a new one determined by Fuzzy-Logic [21]. Katsis et al. employed a correlation feature selection procedure and an AIRS. The application of such an approach can reduce the number of unnecessary biopsies [22]. However, these methods also belong to the supervised machine learning method, which require lots of labeled data.

In order to explore an efficacious method handling and managing auxiliary health problems, and enhance the learning accuracy for the unlabeled breast cancer diagnosis data, a novel automated breast cancer diagnosis algorithm which organically integrates artificial immune with semi-supervised learning (referred as *Aisl*) is proposed. In this way, *Aisl* can utilize the adaptability of immune system and effectively deal with the labeled and unlabeled data. The proposed algorithm is evaluated on two famous UCI breast cancer datasets, and the experimental results show its effectiveness and efficiency.

2. The proposed diagnosis algorithm

Generally, the training process of the proposed algorithm is similar to the response to the invading antigens in mammalian immune system. In a mammalian immune system, the system generates antibody cells, which will respond to an invader, through its presenting characteristics and mutations, these antibody cells develop more and more similar affinity to the antigens. Most antibody cells have a short lifetime, but a small proportion of them become memory cells, which live forever. These memory cells will enable a mammalian immune system to respond rapidly to a next invasion by a previously encountered threat. In the proposed algorithm, the antigens are the training data. When a training example is presented, an initial population of antibody cells is generated and mutated; the resulting antibody cells with the similar affinity to the training data continue to propagate, producing larger numbers of mutated clones or even evolving into memory cells, while those with farther affinity produce fewer offspring and even die out.

From the above descriptions, an individual of breast cancer can be described as an antigen in immune system. The features of antigen represent the attributes of breast cancer. The type of immune cells represents the diagnosis result of breast cancer. Therefore, the breast cancer diagnosis problem can be modeled as a biological immune mechanism.

In order to better describe the algorithm, the definitions of symbols and formulas are given as follows.