Artificial Intelligence in Medicine 52 (2011) 1–9

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

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

A modified artificial immune system based pattern recognition approach—An application to clinical diagnostics

Weixiang Zhao, Cristina E. Davis*

Department of Mechanical and Aerospace Engineering, One Shields Avenue, University of California, Davis, CA 95616, United States

ARTICLE INFO

Article history:
Received 30 March 2010
Received in revised form 2 March 2011
Accepted 11 March 2011

Keywords:
Artificial immune system
Radial basis function
Partial least square regression
Pattern recognition
Breast cancer
Clinical diagnosis

ABSTRACT

Objective: This paper introduces a modified artificial immune system (AIS)-based pattern recognition method to enhance the recognition ability of the existing conventional AIS-based classification approach and demonstrates the superiority of the proposed new AIS-based method via two case studies of breast cancer diagnosis.

Methods and materials: Conventionally, the AIS approach is often coupled with the k nearest neighbor (k-NN) algorithm to form a classification method called AIS-kNN. In this paper we discuss the basic principle and possible problems of this conventional approach, and propose a new approach where AIS is integrated with the radial basis function – partial least square regression (AIS-RBFPLS). Additionally, both the two AIS-based approaches are compared with two classical and powerful machine learning methods, back-propagation neural network (BPNN) and orthogonal radial basis function network (Ortho-RBF network).

Results: The diagnosis results show that: (1) both the AIS-kNN and the AIS-RBFPLS proved to be a good machine leaning method for clinical diagnosis, but the proposed AIS-RBFPLS generated an even lower misclassification ratio, especially in the cases where the conventional AIS-kNN approach generated poor classification results because of possible improper AIS parameters. For example, based upon the AIS memory cells of "replacement threshold = 0.3", the average misclassification ratios of two approaches for study 1 are 3.36% (AIS-RBFPLS) and 9.07% (AIS-kNN), and the misclassification ratios for study 2 are 19.18% (AIS-RBFPLS) and 28.36% (AIS-kNN); (2) the proposed AIS-RBFPLS presented its robustness in terms of the AIS-created memory cells, showing a smaller standard deviation of the results from the multiple trials than AIS-kNN. For example, using the result from the first set of AIS memory cells as an example, the standard deviations of the misclassification ratios for study 1 are 0.45% (AIS-RBFPLS) and 8.71% (AIS-kNN) and those for study 2 are 0.49% (AIS-RBFPLS) and 6.61% (AIS-kNN); and (3) the proposed AIS-RBFPLS classification approaches also yielded better diagnosis results than two classical neural network approaches of BPNN and Ortho-RBF network.

Conclusion: In summary, this paper proposed a new machine learning method for complex systems by integrating the AIS system with RBFPLS. This new method demonstrates its satisfactory effect on classification accuracy for clinical diagnosis, and also indicates its wide potential applications to other diagnosis and detection problems.

1. Introduction

Breast cancer is one of the most common cancers in women, which approximately affects 10% of all women at some stage of their life in the western world [1]. While there are many different diagnostic approaches to detect breast cancer in early stages, the fine needle aspiration (FNA) biopsy method is a reliable and standard test and can be used as a definitive diagnosis method [2]. The information from the extracted biological sample can be examined by pathologists to confirm whether the abnormality is a benign breast disorder or a malignant tumor requiring further testing and treatment. In most cases no single feature or characteristic of the extracted material can separate benign samples from malignant samples [3], and so a reliable diagnosis depends on years of training for a pathologist to identify groups of cancer signature features. As more automated tests are performed, we can consider how computer-aided pattern recognition can increase precision and accuracy of diagnosis by using sophisticated machine learning processes.

An early pattern recognition study of FNA-based breast cancer diagnosis was reported in 1990, where nine cytological characteristics of benign and malignant breast fine-needle aspirates were

---

* Corresponding author. Tel.: +1 530 754 9004; fax: +1 530 752 4158.
E-mail address: cedavis@ucdavis.edu (C.E. Davis).

0933-3657/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.artmed.2011.03.001

employed to establish a classification model for breast cancer diagnosis [3]. The same data set was then used to investigate the diagnosis effect of an instance-based learning strategy [4]. As we can see, with the development of computation and artificial intelligence techniques, a variety of machine learning methods have been developed for clinical diagnosis. For example, discrimination analysis is a typical tool for building multivariate diagnosis models, especially for linear systems [5,6]. Multilayer forward neural networks including the back-propagation neural networks have become a most widely used powerful diagnosis tool for complex nonlinear systems [7,8]. Other approaches such as the fuzzy classifier [9] and self-organization map neural network [10] have also been applied to clinical diagnosis.

Recently, a novel artificial intelligence method termed the artificial immune system (AIS) has been applied to different application areas especially in the pattern recognition field, since it emerged in the 1990s as a bio-inspired computational research tool [11–14]. The main concept of this approach is to use a supervised learning process to create core (representative) data points to represent and cover the sample distribution space of each class. Then, these representative data points will be used to create a model for future prediction. *k* nearest neighbor (*k*-NN) is widely used as a follow-up method to form an AIS-*k*NN classification approach [12,13]. However, the success of this approach depends on the core data points selected by the AIS system. Once the AIS creates memory cells (i.e., core points), the *k*-NN model only uses these points for new sample prediction without further considering any other useful information of the entire existing data, which may result in an unreliable prediction. A detailed principle description of these involved methods will be given in the following sections.

In this paper, we integrate the AIS process with the radial basis function – partial least square regression to form an AIS-RBFPLS approach. By employing two independent clinical diagnosis data sets, we aim to demonstrate the advantage of the proposed AIS-RBFPLS approach over the conventional AIS-*k*NN approach, compare with the two widely used neural network models, and provide suggestions for future applications of our new approach.

2. Material description

The first data set fort this study is the FNA samples to establish a model for separating malignant breast tumor samples from benign samples. The FNA method was used to collect breast tissue materials from patients with a known clinical outcome, and the samples were then mounted on a microscope slide. The samples were then scored by pathologists according to the major cytological characteristics found in the sample. The data for this study was originally from the Clinical Sciences Center at the University of Wisconsin, Madison. The nine cytological characteristics employed for breast cancer detection were: clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, and mitoses. A detailed description of these data can be found at UC Irvine Machine Learning Repository [15]. The sampling process covered two years from 1989 to 1991, in which 699 tissue samples (benign: 458, malignant: 241) were analyzed. The objective for this sample collection and database creation was to establish a diagnosis model using the nine cytological characteristics.

The second data set for this study is the mammography data for breast cancer screening. It contains the patient personal information (age) and three Breast Imaging-Reporting and Data System (BI-RADS) attributes (shape, margin, and density) for 516 benign and 445 malignant masses. The data were originally collected at the Institute of Radiology of the University Erlangen-Nuremberg between 2003 and 2006. The detailed description of this data set can be referred to the above mentioned Machine Learning Repository and literature [16].

3. Method description

3.1. Artificial immune system

Artificial immune systems are a recently developed bio-inspired machine learning method mimicking the response of natural immune systems to pathogen invasion [11]. Basically, when a pathogen invades the human body, special cells circulating within the body called "B-cells" generate antibodies specific to the antigens derived from the pathogen. Each B-cell can only produce one particular antibody. Once a B-cell gets sufficiently stimulated because of the close affinity to a presented antigen, it rapidly produces clones of itself. At the same time, the B-cell surface antigens will mutate to match the antigen as closely as it can, and this process is repeated many times in the human body. After a successful defense against the invading pathogen, a small number of "memory B-cells" stay in the body for sufficiently long periods of time to prevent re-infection. For future pathogen invasions, these memory B-cells can rapidly and efficiently recognize an antigen which is similar to those that the immune system previously fought against. Fig. 1 shows a flowchart of the principle of this learning process. A brief description of the AIS algorithm that effectively mimics this biological process is shown below [12,14].

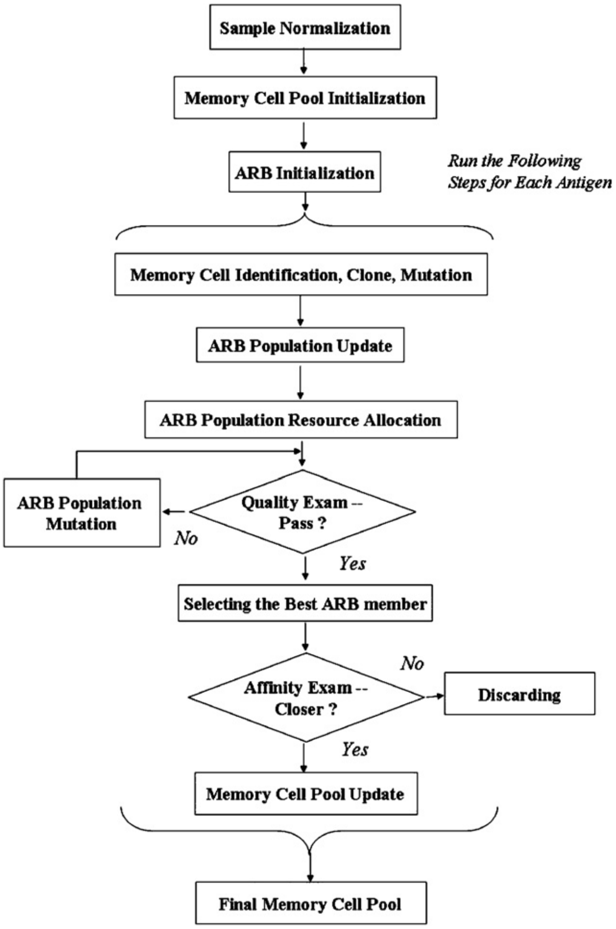

Fig. 1. Flowchart of the AIS learning process.