Available online at www.sciencedirect.com
ScienceDirect
Expert Systems with Applications 34 (2008) 773–779
Expert Systems with Applications
www.elsevier.com/locate/eswa

Computer aided medical diagnosis system based on principal component analysis and artificial immune recognition system classifier algorithm

Kemal Polat, Salih Güneş *

Selcuk University, Department of Electrical and Electronics Engineering, 42075 Konya, Turkey

Abstract

In this study, diagnosis of lung cancer, which is a very common and important disease, was conducted with computer aided medical diagnosis system based on principal component analysis and artificial immune recognition system. The approach system has two stages. In the first stage, dimension of lung cancer dataset that has 57 features is reduced to 4 features using principal component analysis. In the second stage, artificial immune recognition system (AIRS) was our used classifier. We took the lung cancer dataset used in our study from the UCI (from University of California, Department of Information and Computer Science) Machine Learning Database. The obtained classification accuracy of our system was 100% and it was very promising with regard to the other classification applications in literature for this problem.

Keywords: Lung cancer; Principal component analysis (PCA); Artificial immune system; AIRS; Medical diagnosis

1. Introduction

Lung cancers are cancers that begin in the lungs. Other types of cancers may spread to the lungs from other organs. However, these are not lung cancers because they did not start in the lungs. When cancer cells spread from one organ to another, they are called metastases. Research has found several risk factors for lung cancer. A "risk factor" is anything that changes risk of getting a disease. Different risk factors change risk by different amounts.

The risk factors for lung cancer include the following (lungcancer, 2006):

- smoking and being around others' smoke,
- things around us at home or work (such as radon gas),
- personal traits (such as having a family history of lung cancer).

While a new artificial intelligence field named as artificial immune systems (AIS) was emerging in late 1990s, performances of proposed methods were not so good especially for classification problems. However, artificial immune recognition system (AIRS) proposed in 2001 has changed this situation by taking attention among other classifiers with its performance (Watkins, 2001). The reason of its success in classification problems can be found in the following properties of it (Goodman, Boggess, & Watkins, 2003):

- AIRS performs good on very different problems such as large dimensioned feature space problems, problems with many classes, etc.
- AIRS is self-adjusting with regard to its architecture in problem space.

In this study, we have proposed the system that has two stages. Firstly, dimension of lung cancer dataset that has 57 features is reduced to 4 features using principal component analysis. Then, we used artificial immune recognition system (AIRS) classifier diagnosis lung cancer disease.

* Corresponding author. Tel.: +90 332 2232094; fax: +90 332 2410635.
E-mail addresses: kpolat@selcuk.edu.tr (K. Polat), sgunes@selcuk.edu.tr (S. Güneş).

0957-4174/$ - see front matter
doi:10.1016/j.eswa.2006.10.011

The obtained classification accuracy of our system was 100% and it was very promising with regard to the other classification applications in Literature for this problem.

The rest of the paper is organized as follows. Section 2 gives the background including the used lung cancer dataset and natural and artificial immune system. In each subsection of this section, the detailed information was given. We give the proposed approach in Section 3. In Section 4, we give the experimental data to show the effectiveness of our method. This section also includes the discussion of these results in specific and general manner. Consequently in Section 5, we conclude the paper with summarization of results by emphasizing the importance of this study and mentioning about some future work.

2. Background

2.1. The lung cancer dataset

The dataset used in this study was obtained from the archives of machine learning datasets at the University of California, Irvine (Blake & Merz, 1996). This data was first used by Hong and Young to illustrate the power of the optimal discriminant plane even in ill-posed settings. The purpose of the dataset is to predict the one from three classes of lung cancer given the results of various medical tests carried out on a patient. . This database contains 57 attributes, which have been extracted from a larger set. The database originally contained 32 examples. There are three classes. There are missing values in the dataset. In such cases, probabilistic values were assigned according to the distribution of the known values for the attributes. Class distribution of lung cancer dataset contains 9 examples from class1, 13 examples from class 2 and 10 examples from class 3.

2.2. Artificial immune system

The natural immune system is a distributed novel-pattern detection system with several functional components positioned in strategic locations throughout the body. Immune system regulates defence mechanism of the body by means of innate and adaptive immune responses. Between these, adaptive immune response is more important for us because it contains metaphors like recognition, memory acquisition, diversity, self-regulation, etc. The main architects of adaptive immune response are lymphocytes, which divide into two classes as *T* and *B* lymphocytes (cells), each having its own function. Especially *B* cells have a great importance because of their secreted antibodies (Abs) that take very critical roles in adaptive immune response.

The simplified working procedure of our immune system is illustrated in Fig. 1. Specialized antigen presenting cells (APCs) called macrophages circulates through the body and if they encounter an Antigen, they ingest and fragment them into antigenic peptides (I). The pieces of these peptides are displayed on the cell surface by major histocompatibility complex (MHC) molecules existing in the digesting APC. The presented MHC-peptide combination on the cell surface is recognized by the *T*-cells causing them to be activated (II). Activated *T* cells secrete some chemicals as alert signals to other units in response to this recognition (III). *B* cells, one of the units that take these signals from the *T* cells become activated with the recognition of Antigen by their Antibodies occurred in the same time (IV). When activated, *B* cells turn into plasma cells that secrete bound Antibodies on their surfaces (V). Secreted Antibodies bind the existing Antigens and neutralize them signalling other components of immune system to destruct the antigen-antibody complex (VI) (De Castro & Timmis, 2002). For detailed information about immune system refer to (Abbas & Lichtman, 2003).

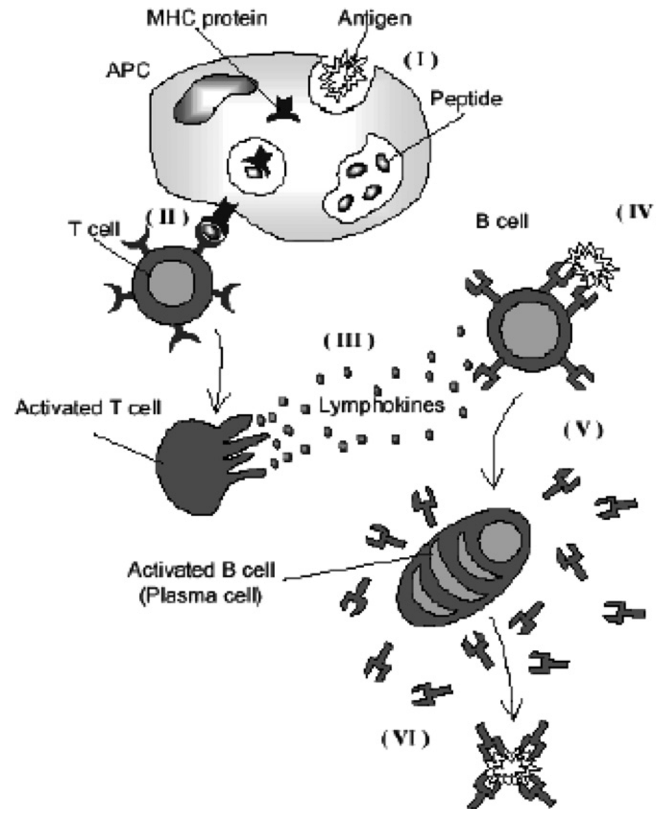

Fig. 1. Immune response.

Artificial immune systems (AISs) emerged in the 1990s as a new computational research area. Artificial immune systems link several emerging computational fields inspired by biological behaviour such as artificial neural networks and artificial life.

In the studies conducted in the field of AIS, *B* cell modelling is the most encountered representation type. Different representation methods have been proposed in that modelling. Among these, shape-space representation is the most commonly used one (Perelson & Oster, 1979).

The shape-space model (*S*) aims at quantitatively describing the interactions among antigens (Ags), the