Available online at www.sciencedirect.com
ScienceDirect
Expert Systems with Applications 34 (2008) 214–221
Expert Systems with Applications
www.elsevier.com/locate/eswa

Principles component analysis, fuzzy weighting pre-processing and artificial immune recognition system based diagnostic system for diagnosis of lung cancer

Kemal Polat *, Salih Güneş

Selcuk University, Department of Electrical and Electronics Engineering, 42035 Konya, Turkey

Abstract

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. It is evident that usage of machine learning methods in disease diagnosis has been increasing gradually. In this study, diagnosis of lung cancer, which is a very common and important disease, was conducted with such a machine learning system. In this study, we have detected on lung cancer using principles component analysis (PCA), fuzzy weighting pre-processing and artificial immune recognition system (AIRS). The approach system has three stages. First, dimension of lung cancer dataset that has 57 features is reduced to four features using principles component analysis. Second, a new weighting scheme based on fuzzy weighting pre-processing was utilized as a pre-processing step before the main classifier. Third, artificial immune recognition system was our used classifier. We took the lung cancer dataset used in our study from the UCI 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: Principles component analysis; Artificial immune system; AIRS; Fuzzy weighting pre-processing; Lung cancer; 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 (http://www.cdc.gov/lungcancer/basic_info/index.htm):

- 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).

Having so many factors to analyze to diagnose the lung cancer of a patient makes the physician's job difficult. A physician usually makes decisions by evaluating the current test results of a patient and by referring to the previous decisions she made on other patience with the same condition. The former method depends strongly on the physician's knowledge. On the other hand, the latter depends on the physician's experience to compare her patient with her earlier patients. This job is not easy considering the number of factors she has to evaluate. In this crucial step, she may need an accurate tool that lists her previous decisions on the patient having same (or close to same) factors.

* Corresponding author. Tel.: +90 332 223 2056; fax: +90 332 241 0635.
E-mail addresses: kpolat@selcuk.edu.tr (K. Polat), sgunes@selcuk.edu.tr (S. Güneş).

0957-4174/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2006.09.001

Motivated by the need of such an important classification method, in this study, we propose a method to diagnose the lung cancer. The proposed method uses PCA, AIRS classification system and a weighting process, which is based on fuzzy weighted pre-processing method. First, dimension of lung cancer dataset that has 57 features is reduced to 4 features using principles component analysis. Second, a new weighting scheme based on fuzzy weighting pre-processing was utilized as a pre-processing step before the main classifier. Third, artificial immune recognition system was our used classifier. We took the lung cancer dataset used in our study from the UCI 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.

The remaining of the paper is organized as follows. In Section 2, we give a brief background on natural and artificial immune systems. We present the proposed method in Section 3. In Section 4, we give the experimental data to show the effectiveness of our method. Finally, we conclude this paper in Section 5 with future directions.

2. Natural and artificial immune systems

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 defense mechanism of the body by means of innate and adaptive immune responses. Between these, adaptive immune response is much 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 can be divided 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 the human 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 signaling other components of immune system to destruct the antigen–antibody complex (VI) (De Castro & Timmis, 2002). For more detailed information about immune system refer to Abbas and Lichtman (2003), Şahan, Kodaz, Güneş, and Polat (2004), Perelson and Oster (1979).

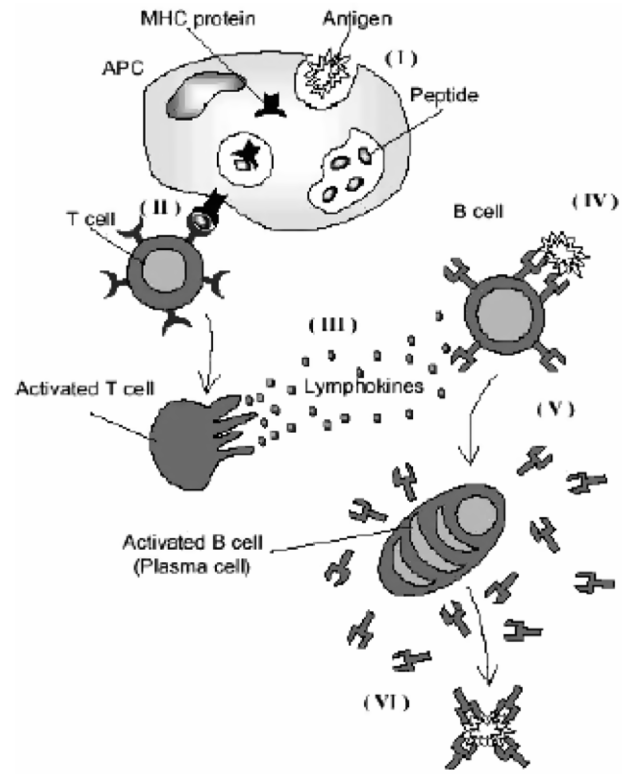

Fig. 1. Immune response.

Artificial immune systems emerged in the 1990s as a new computational research area. Artificial immune systems link several emerging computational fields inspired by biological behavior such as artificial neural networks and artificial life (De Castro & Timmis, 2002).

In the studies conducted in the field of AIS, B cell modeling is the most encountered representation type. Different representation methods have been proposed in that modeling. Among these, shape-space representation is the most commonly used one (De Castro & Timmis, 2002).

The shape-space model (S) aims at quantitatively describing the interactions among antigens (Ags), the foreign elements that enter the body like microbe...etc., and antibodies (Ag–Ab). The set of features that characterize a molecule is called its *generalized shape*. The Ag–Ab representation (binary or real-valued) determines a distance measure to be used to calculate the degree of interaction between these molecules. Mathematically, the generalized shape of a molecule (m), either an antibody or an antigen, can be represented by a set of coordinates $m = \langle m_1, m_2, \ldots, m_L \rangle$, which can be regarded as a point in an L-dimensional real-valued shape-space ($m \in S^L$). In this work, we used real strings to represent the molecules. Antigens and antibodies are considered to be the same length L.