Available online at www.sciencedirect.com
ScienceDirect
Expert Systems with Applications 32 (2007) 625–631
Expert Systems with Applications
www.elsevier.com/locate/eswa

Automatic detection of heart disease using an artificial immune recognition system (AIRS) with fuzzy resource allocation mechanism and *k*-nn (nearest neighbour) based weighting preprocessing

Kemal Polat *, Seral Şahan, Salih Güneş

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

Abstract

It is evident that usage of machine learning methods in disease diagnosis has been increasing gradually. In this study, diagnosis of heart disease, which is a very common and important disease, was conducted with such a machine learning system. In this system, a new weighting scheme based on *k*-nearest neighbour (*k*-nn) method was utilized as a preprocessing step before the main classifier. Artificial immune recognition system (AIRS) with fuzzy resource allocation mechanism was our used classifier. We took the dataset used in our study from the UCI Machine Learning Database. The obtained classification accuracy of our system was 87% and it was very promising with regard to the other classification applications in the literature for this problem.

Keywords: Heart disease; Artificial immune system; AIRS; *k*-nn based weighting preprocessing; Expert systems

1. Introduction

One of the central problems of the information age is dealing with the enormous amount of raw information that is available. More and more data is being collected and stored in databases or spreadsheets. As the volume increases, the gap between generating and collecting the data and actually being able to understand it is widening. In order to bridge this knowledge gap, a variety of techniques known as data mining or knowledge discovery is being developed. Knowledge discovery can be defined as the extraction of implicit, previously unknown, and potentially useful information from real world data, and communicating the discovered knowledge to people in an understandable way (Fayyad, Piatetsky-Shapiro, Smyth, & Uthurusamy, 1996; Michie, 1991; Piatetsky-Shapiro & Frawley, 1991).

Heart disease is any disorder that affects the heart's ability to function normally. The most common cause of heart disease is narrowing or blockage of the coronary arteries, which supply blood to the heart itself. This happens slowly over time.[1]

Extensive clinical and statistical studies have identified several factors that increase the risk of coronary heart disease and heart attack. Major risk factors are those which research has shown to significantly increase the risk of heart and blood vessel (cardiovascular) disease. Other factors are associated with increased risk of cardiovascular disease, but their significance and prevalence have not yet been precisely determined. They are called contributing risk factors. The American Heart Association has identified several risk factors. Some of them can be modified, treated or controlled, and some cannot. The more risk factors you have, the greater your chance of developing coronary heart

* Corresponding author. Tel.: +90 332 2232098; fax: +90 332 2410635.
E-mail addresses: kpolat@selcuk.edu.tr (K. Polat), seral@selcuk.edu.tr (S. Şahan), sgunes@selcuk.edu.tr (S. Güneş).

[1] http://health.allrefer.com/health/heart-disease-info.html (last accessed: 9 October 2005).

0957-4174/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2006.01.027

disease. Also, the greater the level of each risk factor, the greater the risk.[2]

In this study, heart disease was diagnosed by using fuzzy-AIRS classification system in which a weighting process based on k-nn method was used as a preprocessing step. While conducting this study, we firstly applied the k-nn based weighting process to the dataset and weighted it in the interval $[0,1]$. After this preprocessing step, the weighted dataset was presented to the main fuzzy-AIRS classifier algorithm. The obtained classification accuracy was found to be 87.00%.

The rest of the paper is organized as follows. Section 2 gives the background information including the heart disease classification problem, previous research in corresponding area and a brief introduction to natural and artificial immune systems. We explained the method in Section 3 with subtitles of the proposed method and measures for performance evaluation. In each subsection of that section, the detailed information is given. The results obtained in applications are given in Section 4. 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. Heart disease classification problem

This database comes from the Cleveland Clinic Foundation and was supplied by Robert Detrano, M.D., Ph.D. of the V.A. Medical Center, Long Beach, CA. It is part of the collection of databases at the University of California, Irvine collected by David Aha. The purpose of the dataset is to predict the presence or absence of heart disease given the results of various medical tests carried out on a patient. This database contains 13 attributes, which have been extracted from a larger set of 75. The database originally contained 303 examples but 6 of these contained missing class values and so were discarded leaving 297.27 of these were retained in case of dispute, leaving a final total of 270. There are two classes, presence and absence (of heart disease). This is a reduction of the number of classes.[3]

2.2. Previous research

As for other clinical diagnosis problems, classification systems have been used for heart disease diagnosis problem, too. When the studies in the literature related with this classification application are examined, it can be seen that a great variety of methods were used which reached high classification accuracies using the dataset taken from UCI machine learning repository. Among these, ToolDiag, RA obtained 50.00% classification accuracy by using IB1-4 algorithm.[4] WEKA, RA obtained a classification accuracy of 58.50% using InductH algorithm while ToolDiag, RA reached to 60.00% with RBF algorithm.[4] Again, WEKA, RA applied FOIL algorithm to the problem and obtained a classification accuracy of 64.00%.[4] MLP+BP algorithm that was used by ToolDiag, RA reached to 65.60%.[4] The classification accuracies obtained with T2, 1R, IB1c and K* which were applied by WEKA, RA are 68.10%, 71.40%, 74.00% and 76.70%, respectively.[4] In addition, Robert Detrano used logistic regression algorithm and obtained 77.0% classification accuracy. Moreover, Cheung utilized C4.5, Naive Bayes, BNND and BNNF algorithms and reached the classification accuracies 81.11%, 81.48%, 81.11% and 80.96%, respectively (Cheung, 2001). Higher classification accuracy was obtained by WEKA, RA using Naive–Bayes algorithm as 83.60%.[4] Among these studies, AIRS, which is our base algorithm, was also used and reached a classification accuracy of 84.50% (Polat, Şahan, Kodaz, & Güneş, 2005).

2.3. 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. The 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 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 at the same time (IV).

[2] http://www.americanheart.org/presenter.jhtml?identifier=4726 (last accessed: 9 October 2005).

[3] ftp://ftp.ics.uci.edu/pub/machine-learning-databases (last accessed: 4 October 2005).

[4] http://www.phys.uni.torun.pl/kmk/proiects/datasets.html#Sheart (last accessed: 9 October 2005).