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

A hybrid approach to medical decision support systems: Combining feature selection, fuzzy weighted pre-processing and AIRS

*Kemal Polat**, *Salih Güneş*

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

ARTICLE INFO

Article history:
Received 14 January 2007
Received in revised form
21 July 2007
Accepted 31 July 2007

Keywords:
AIRS
Fuzzy weighted pre-processing
Feature selection
Heart disease
Hepatitis disease
Medical decision-making

ABSTRACT

This paper presents a hybrid approach based on feature selection, fuzzy weighted pre-processing and artificial immune recognition system (AIRS) to medical decision support systems. We have used the heart disease and hepatitis disease datasets taken from UCI machine learning database as medical dataset. Artificial immune recognition system has shown an effective performance on several problems such as machine learning benchmark problems and medical classification problems like breast cancer, diabetes, and liver disorders classification. The proposed approach consists of three stages. In the first stage, the dimensions of heart disease and hepatitis disease datasets are reduced to 9 from 13 and 19 in the feature selection (FS) sub-program by means of C4.5 decision tree algorithm (CBA program), respectively. In the second stage, heart disease and hepatitis disease datasets are normalized in the range of [0,1] and are weighted via fuzzy weighted pre-processing. In the third stage, weighted input values obtained from fuzzy weighted pre-processing are classified using AIRS classifier system. The obtained classification accuracies of our system are 92.59% and 81.82% using 50–50% training-test split for heart disease and hepatitis disease datasets, respectively. With these results, the proposed method can be used in medical decision support systems.

1. Introduction

In this article, we have proposed a new hybrid approach to medical decision support systems, which help to physicians related with pre-decision about any disease. The proposed method includes three parts: (i) dimensionality reduction using feature selection algorithm, (ii) data-weighted process using fuzzy weighted pre-processing, and (iii) classification process using AIRS classifier algorithm. We have assessed on the two medical datasets including heart disease and hepatitis disease datasets taken from UCI machine learning database.

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 significantly increase the risk of heart and blood vessel (cardiovascular) disease. Other factors are associated with increased risk of cardiovascular disease, but their signifi-

* *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).
0169-2607/$ – see front matter © 2007 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.cmpb.2007.07.013

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

Most of the time the hepatitis diagnoses is made by a routine blood testing or during a blood donation. The hepatitis is a viral infection that also was transmitted by blood or blood products in the past, when there was no test available to screen for this infection. Risk factors are as follows: blood transfusions, tatoos and piercing, drug abuse, haemodialysis, health workers, and sexual contact with hepatitis carrier [3].

With improvements in medical knowledge systems in medical institutes and hospitals, determining useful knowledge is becoming more difficult. Especially, because the conventional manual data analysis techniques are not effective in diagnosis, using computer-based analyses are becoming inevitable in disease diagnosis. So, it is the time to develop modern, effective and efficient computer-based systems for decision support. There are a number of data analysis techniques: statistical, machine learning and data abstraction. Medical analysis using machine learning techniques has begun to be conducted for last 20 years. The advantages of using machine learning schemes in medical analysis have caused human support and costs to decrease and caused diagnosis accuracy to increase [4].

The use of classifier systems in medical diagnosis is increasing gradually. There is no doubt that evaluation of data taken from patient and decisions of experts are the most important factors in diagnosis. But, expert systems and different artificial intelligence techniques for classification also help experts in a great deal. Classification systems, helping possible errors that can be done because of fatigued or inexperienced expert to be minimized, provide medical data to be examined in shorter time and in more detail.

In this study, the proposed approach consists of three stages. In the first stage, the dimensions of heart disease and hepatitis disease datasets are reduced to 9 from 13 and 19 in the feature selection (FS) sub-program by means of C4.5 decision tree algorithm (CBA program), respectively. In the second stage, heart disease and hepatitis disease datasets are normalized in the range of [0,1] and are weighted via fuzzy weighted pre-processing. In the third stage, weighted input values obtained from fuzzy weighted pre-processing are classified using AIRS classifier system. The performance of the proposed method was tested with regard to the classification accuracy, sensitivity and specificity values, and confusion matrix. The obtained classification accuracies of our system are 92.59% and 81.82% using 50–50% training-test split for heart disease and hepatitis disease datasets, respectively. With these results, the proposed method can be used in medical decision support systems.

The rest of the paper is organized as follows. Section 2 gives the used medical datasets information including heart disease and hepatitis disease datasets classification problem, previous research in corresponding area. We explained the method in Section 3 with subtitles of proposed new medical diagnosis method. 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. Used medical datasets

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. The database originally contained 303 examples but 6 of these contained missing class values and so were discarded leaving 297. Twenty-seven 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 [5]. Number of presence (normal) is 120 subjects and number of absence (of heart disease) is 150 subjects.

Attributes of symptoms that are obtained from patients are as follows [5]:

1. age: (in the range of 29 and 77)
2. sex: male, female
3. chest pain type (4 values): 1, 2, 3, 4
4. resting blood pressure: (in the range of 94 and 200)
5. serum cholesterol in mg/dl (in the range of 126 and 564)
6. fasting blood sugar >120 mg/dl (in the range of 126 and 564)
7. resting electrocardiograph results (values 0, 1, 2)
8. maximum heart rate achieved (in the range of 71 and 202)
9. exercise-induced angina (either 0 or 1)
10. old peak = ST depression induced by exercise relative to rest (in the range of 0 and 6.2)
11. the slope of the peak exercise ST segment: 1, 2, and 3
12. number of major vessels (0–3) colored by fluoroscopy: 0–3
13. thal: 3 = normal; 6 = fixed defect; 7 = reverse defect.

2.2. *Previous research on the heart disease*

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 [6]. WEKA & Ra obtained a classification accuracy of 58.50% using InductH algorithm while ToolDiag & RA reached to 60.00% with RBF algorithm [6]. Again, WEKA & Ra applied Foil algorithm to the problem and obtained a classification accuracy of