Expert Systems with Applications 40 (2013) 96–104

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

Computational intelligence for heart disease diagnosis: A medical knowledge driven approach

Jesmin Nahar [a,*], Tasadduq Imam [a], Kevin S. Tickle [a], Yi-Ping Phoebe Chen [b]

[a] *Faculty of Arts, Business, Informatics and Education, Central Queensland University, Queensland, Australia*
[b] *Department of Computer Science and Computer Engineering, La Trobe University, Melbourne, Victoria 3086, Australia*

ARTICLE INFO

Keywords:
Cleveland data
Heart disease
Computational intelligence
Classification
Feature selection

ABSTRACT

This paper investigates a number of computational intelligence techniques in the detection of heart disease. Particularly, comparison of six well known classifiers for the well used Cleveland data is performed. Further, this paper highlights the potential of an expert judgment based (i.e., medical knowledge driven) feature selection process (termed as MFS), and compare against the generally employed computational intelligence based feature selection mechanism. Also, this article recognizes that the publicly available Cleveland data becomes imbalanced when considering binary classification. Performance of classifiers, and also the potential of MFS are investigated considering this imbalanced data issue. The experimental results demonstrate that the use of MFS noticeably improved the performance, especially in terms of accuracy, for most of the classifiers considered and for majority of the datasets (generated by converting the Cleveland dataset for binary classification). MFS combined with the computerized feature selection process (CFS) has also been investigated and showed encouraging results particularly for NaiveBayes, IBK and SMO. In summary, the medical knowledge based feature selection method has shown promise for use in heart disease diagnostics.

1. Introduction

Various classification and regression processes have been used to identify heart disease (Boors et al., 2000; Das, Turkoglu, & Sengur, 2009; Detrano et al., 1989; El-hanjouri, Alkhaldi, Hamdy, & Alim, 2002; Skalak, 1997). In particular, focus has been made on the University of California Irvine (UCI) heart disease dataset (also known as the Cleveland dataset (Uci, 2009) and different computational intelligence algorithms have been used. But, existing investigations are, to the best of the author's knowledge, yet to show a comparative research that considers modern classification techniques and imbalanced nature of the data, and employs a feature selection process which incorporates medical knowledge. Medical knowledge is important for feature selection in this area since computer automated process may remove features important or select feature that are less likely related from clinical view. The research presented in this paper highlights this issue and the findings may aid in contributing to future identification of heart disease.

The plan of this paper is as follows: Section 2 provides an overview of existing research on using computational intelligence techniques in heart diseases diagnosis. Section 3 details the datasets and 4 demonstrate the experimental setup that has been used in this research. Section 5 then presents the comparative research of the different classification algorithms and describes the best suited ones for this problem. Section 6 describes significant risk factors for heart disease from medical point of view. Section 7 presents the results of the comparison of computer feature selection (CFS) process and medical knowledge based feature selection for heart disease dataset. Section 8 proposes the medical knowledge motivated feature selection (MFS), as well as a process combining CFS with MFS. Finally, Section 9 concludes the paper with a summary of findings and future research directions.

2. Computational intelligence for heart disease diagnostics

This section provides an overview of existing research on using computational intelligence techniques in heart diseases diagnosis and points to the limitations that motivated this research. Cardiovascular disease is a highly mortal disease with over 17 million deaths globally (Smith, 2010). So, early detection and treatment of the disease are imperative. Researchers have used different computational intelligence techniques to improve heart disease diagnostics over the years. A particular heart disease diagnostic dataset widely popular with data mining researchers is the publicly available University of California Irvine, Cleveland dataset (Uci, 2009). Some of the key researches on this datasets are:

* Corresponding author. Tel.: +61 07 40232112; fax: +61 07 49309700.
E-mail addresses: j.nahar@cqu.edu.au (J. Nahar), t.imam@cqu.edu.au (T. Imam), k.tickle@cqu.edu.au (K.S. Tickle), phoebe.chen@latrobe.edu.au (Y-.P.P. Chen).

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2012.07.032

- Aha & Kibler (1988) used the dataset to predict effectiveness of instance-based algorithms and achieved 77% and 74.8% accuracy for NTgrowth and C4.5 techniques.
- Detrano et al. (1989) investigated a probabilistic algorithm to diagnose the risk of coronary artery disease and concluded that patients experiencing chest pain and transitional disease occurrences are the higher risk subjects.
- Gennari, Langley, & Fisher (1989) explored a conceptual clustering system and gained an acceptable accuracy (78.9%).
- Edmonds (2005) worked on the Cleveland data set with focus on comparing global evolutionary computation approaches, and observed some prediction performance improvement with a new approach. However, performance of the proposed technique is dependent on the attributes selected by the algorithm.

Other than these works, several researches have focused on diverse aspects of heart disease diagnosis on different datasets (Avci, 2009; Boors et al., 2000; Doyle, Temko, Marnane, Lightbody, & Boylan, 2010; El-hanjouri et al., 2002; Gamboa, Mendoza, Orozco, VARGAS, & Gress, 2006; Maglogiannis, Loukis, Zafiropoulos, & Stasis, 2009; Obayya & Abou-chadi, 2008; Zheng, Jiang, & Yan, 2006; Kim, Lee, Cho, & Oh, 2008). Also, different researchers have used different machine learning techniques in related research. These include: fuzzy support vector clustering for the identification of heart disease (Gamboa et al., 2006), prototype development using data mining techniques, mainly decision trees, Naive Bayes and Neural Networks, (Palaniappan & Awang, 2008) diagnostic system improved using feature extraction and Hidden Markov Models (HMM) (El-hanjouri et al., 2002), a data fusion approach recommended for classifying heart diseases (Obayya & Abou-chadi, 2008), an intelligent system based on genetic-support vector machines (GSVM) (Avci, 2009), use of an automated detection system based on the SVM classification (Maglogiannis et al., 2009), a committee machine (CM) based on an ensemble of Multilayer Perceptions (MLP) (Zheng et al., 2006), a computerized cardiovascular disease diagnosis and categorization system (Kim et al., 2008) and decision trees and SVM to predict heart disease (Soman, Shyam, & Madhavdas, 2003).

Feature selection has also been applied in heart disease diagnostics, but for mainly datasets other than Cleveland. For instance, Zhao, Chen, Hou, Zheng, & Wang (2010) used backward elimination procedure along with a novel algorithm, Fan & Chaovalitwongse (2010) suggested a novel optimization framework for getting improved feature selection in classification. Several other researchers have also noted impact of feature selection in different heart disease diagnosis (Chang, 2010; Hanbay, 2009; Qazi et al., 2007; Zhao, Guo et al., 2010). Further, feature selection processes have often been found to improve the prediction performance of different classifiers (Abraham, Simha, & Iyengar, 2007; Cheng, Wei, & Tseng, 2006; Devaney & Ram, 1997; Polat & Guenes, 2009; Sethi & Jain, 2010; Wang & Ma, 2009; Zhao, Chen et al., 2010).

It is observed that a number of different classifier have been used to diagnose heart disease in the different studies. The comparison of different algorithms in order to identify the heart disease, however, has to date not received appropriate focus. In addition, the literature has not taken into account medical knowledge based feature selection for medical datasets during the classification of heart disease. Computer based feature selection (CFS) selects features randomly, through calculating the significance of the attributes and by considering the individual predictive capacity. So, there is a chance to discard medically important factors for a specific disease. For instance, as shown in Fig. 4, applying computerized feature selection (CFS) on Cleveland dataset (with healthy as the positive class) discards medically established attributes like age, cholesterol, fasting blood sugar, resting blood pressure and ECG characteristics. This sort of outcomes is doubted by medical practitioners and reduces the significance of the automated system. So, a feature selection process motivated by medical knowledge is important.

The literature also indicates that in most cases complex and time intensive algorithms have been recommended. Well-known standard classification algorithms are, however, more easily accessible due to its availability in different software packages. From the medical practitioner's point of view, in particular, a comprehensive analysis of well-established classifiers is, so, of essence.

This research focuses on these issues. As Cleveland dataset is considered a benchmark data in many existing research, this research also uses this dataset. The study provides a comparative suitability of commonly used classifiers. In addition, the research investigates medical knowledge guided feature selection process for classification of heart disease.

3. Dataset details

As mentioned earlier, the popular and publicly available UCI heart disease dataset is used in this research. The UCI heart disease dataset consists of a total 76 attributes. However, majority of the existing studies have used only a maximum of 14 attributes (Uci, 2009; Uci, 2010). Different datasets have been based on the UCI heart disease data. Computational intelligence researchers, however, have mainly used the Cleveland dataset consisting of 14 attributes. The 14 attributes of the Cleveland dataset along with the values and data types are as follow (Uci, 2009; Uci, 2010).

1. Age: age in years (*numeric*);
2. Sex: male, female (*nominal*);
3. Chest pain type (CP): (a) typical angina (angina), (b) atypical angina (abnang), (c) non-anginal pain (notang), (d) asymptomatic (asympt) (*nominal*). From medical point of view,
   (a) Typical angina is the condition in which the past history of the patient shows the usual symptoms and so the possibility of having coronary artery blockages is high (Baliga & Eagle, 2008; Diagnosis, 2010; Kaul, 2010).
   (b) Atypical angina refers to the condition that the patient's symptoms are not detailed and so the probability of blockages is lower (Baliga & Eagle, 2008; Diagnosis, 2010; Kaul, 2010).
   (c) Non-angina pain is the stabbing or knife-like, prolonged, dull, or painful condition that can last for short or long periods of time (Diagnosis, 2010; Mengel & Schwiebert, 2005; Society, 1945).
   (d) Asymptomatic pain shows no symptoms of illness or disease and possibly will not cause or exhibit disease symptoms (Pickett, 2000; Freedc, 2010);
4. Trestbps: patient's resting blood pressure in mm Hg at the time of admission to the hospital (*numeric*);
5. Chol: Serum cholesterol in mg/dl;
6. Fbs: Boolean measure indicating whether fasting blood sugar is greater than 120 mg/dl: (1 = True; 0 = false) (*nominal*);
7. Restecg: electrocardiographic results during rest. Three types of values normal (norm), abnormal (abn): having ST-T wave abnormality, ventricular hypertrophy (hyp) (*nominal*);
8. Thalach: maximum heart rate attained (*numeric*);
9. Exang: Boolean measure indicating whether exercise induced angina has occurred: 1 = yes, 0 = no (*nominal*);
10. Oldpeak: ST depression brought about by exercise relative to rest (*numeric*);
11. Slope: the slope of the ST segment for peak exercise. Three types of values upsloping, flat, downsloping (*nominal*);