Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

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

Technical Note

Early detection of liver disease using data visualisation and classification method

Xiaofeng Zhou[a], Yonglai Zhang[a,b,c,*], Mingrui Shi[d], Haibo Shi[a], Zeyu Zheng[a,e]

[a] Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang, China
[b] University of Chinese Academy of Sciences, Beijing, China
[c] Department of Economics and Management, Liaoning Technical University, Fuxin, China
[d] Department of Control Science and Engineering, Zhejiang University, Hangzhou, China
[e] Department of Physics and Centre for Computational Science and Engineering, National University of Singapore, Singapore

ARTICLE INFO

Article history:
Received 14 October 2013
Received in revised form 29 January 2014
Accepted 17 February 2014
Available online 15 March 2014

Keywords:
Machine learning
Liver disease
Classification
SVDD
Visualisation
GSO

ABSTRACT

Detection of early-stage liver diseases is a challenge in medical field. Automated diagnostics based on machine learning therefore could be very important for liver tests of patients. This paper investigates 225 liver function test records (each record include 14 features), which is a subset from 1000 patients' liver function test records that include the records of 25 patients with liver disease from a community hospital. We combine support vector data description (SVDD) with data visualisation techniques and the glowworm swarm optimisation (GSO) algorithm to improve diagnostic accuracy. The results show that the proposed method can achieve 96% sensitivity, 86.28% specificity, and 84.28% accuracy. The new method is thus well-suited for diagnosing early liver disease.

1. Introduction

Liver disease is one of the most common diseases. Most of the risk factors for liver disease and liver cancer are discussed in [1]. Worldwide, Liver cancer is the fifth most commonly diagnosed cancer in men and the seventh in women [2].

According to Department of Health statistics, liver disease has consistently been among the top 10 fatal diseases in Beijing. In 2010, liver cancer accounted for 11.3% of cancer mortality in men and is second only to lung cancer. Symptoms are easily overlooked in the initial stage, and by the time they appear, the patient has missed the best opportunity for treatment. Effective early diagnosis is therefore of paramount importance, and automation of this diagnosis is highly desirable. Much effort has been devoted to solving this problem.

Automated diagnostics for liver disease have been intensively researched in recent years, and many methods have been proposed and applied [3,4,1,5–8]. Data mining (DM) techniques, such as artificial neural networks (ANN), intelligent algorithms, fuzzy sets, and support vector machines (SVMs), have recently found medical application. Most of this work falls into two broad categories: diagnostics for patients and early detection for the general population.

On the one hand, Nakano were the first to propose a machine learning technique in which artificial neural networks differentiate between the two subtypes of chronic active hepatitis, mild and severe, based on five blood biochemical parameters [3]. The ANN network correctly diagnosed 78% of the cross validation control group with data from 31 patients. Onisko used a probabilistic causal model that included 16 liver disorders and 40 features in her initial report of the diagnostic performance test results [4]. In the long term, the model may be unsuitable if it misses major disorders that require immediate attention, such as liver cancer. Growth curve analysis using logistic regression, decision tree, and neural networks showed that liver disease was accurately diagnosed in 72.55% of the cases and that the sensitivity was 78.62% when a neural network was used [1]. In the pattern recognition literature, Comak presented a hybrid method based on combining least square support vector machines (LSSVM) with fuzzy weighting preprocessing to diagnose liver disorders using the standard BUPA Liver Disorders Dataset benchmark [5]. The highest classification accuracy yet reported for this dataset is 94.29%. Based on the same dataset, Polat proposed a classification algorithm based

* Corresponding author at: No.19, Feiyun Road, Hunnan New District, Shenyang City 110179, China. Tel.: +86 024 83601301; fax: +86 024 83602708.
E-mail address: zhangyonglai@sia.cn (Y. Zhang).

http://dx.doi.org/10.1016/j.bspc.2014.02.006
1746-8094/

on a fuzzy artificial immune recognition system (AIRS) for liver diagnostics [6]. The accuracy was 83.36%, but the algorithm gained an advantage by means of a shorter classification time. In-depth research was carried out both for patient diagnosis and for early prevention in the general population.

On the other hand, Lin applied CART (classification and regression tree) and CBR (case-based reasoning) techniques to the diagnosis of early stage liver disease [7], achieving a diagnostic accuracy of 90%. Lin further constructed an intelligent liver disease diagnostic model capable of distinguishing among the various types of liver disease [8]. Ramana evaluated four classification algorithms, which include a naïve Bayesian classifier, C4.5 (the decision tree learner), back-propagation neural network, and SVMs, for the classification of two liver patient datasets based on four criteria [9]. The results of an exploratory analysis of USA and INDIA liver disease datasets were reported in [10].

Most previous research on the development of liver disease diagnosis models uses datasets in which the normal and abnormal samples are comparable in sizes. However, in early diagnosis for realistic populations, we often need to solve the classification problem when the normal samples greatly outnumber the abnormal ones. By data visualisation, we refer to the reduction of number of features (14 in our dataset) into 2 or 3 dimensions, so that the samples can be plotted on a graphical representation. In such a manner, the two classes of samples (normal and abnormal) can be visually distinguished. Data visualisation combined with traditional machine learning plays a central role in an intelligent diagnostic model. This model requires an inter-play between automated analysis and human judgement.

Principal component analysis (PCA) and Independent Component Analysis (ICA) are linear methods for dimensionality reduction. Manifold learning is a nonlinear dimensionality reduction approach. Locally linear embedding (LLE) and isometric feature mapping (ISOMAP) have been frequently applied for dimensionality reduction. Recently, various dimensionality reduction methods in the machine learning were reported such as self-organizing maps (SOMs), Hessian LLE, diffusion maps, Laplacian eigenmaps, and others [11–15]. Visualisation methods that reflect this inherent structure to support the user during the process of dimensionality reduction are therefore becoming extensively used in more and more fields [16–18]. Nevertheless, data visualisation has yet to find wide application in the medical field.

The present paper presents a data visualisation and classification method for early-stage diagnosis of liver disease in the general population based on a realistic dataset. A dimensionality reduction technique is first applied to visualise the essential signals after a reduction to two or three dimensions. The method then employs a classification algorithm to distinguish between healthy and diseased livers based on a realistic dataset used for machine learning in medical diagnostics.

This paper is organised as follows. Section 2 describes our clinical dataset, the dimensionality reduction methods used for visualisation, and the diagnostic methods. Section 3 describes the two steps in liver disease diagnostics: visualisation (Section 3.1) and actual diagnosis (Section 3.2). Section 4 has the discussion. In this section, the new method is applied to a realistic and compared with some other liver diagnostic methods. Section 5 summarises our conclusions.

2. Materials and methods

2.1. Clinical data

Data from liver function tests for 1000 patients in a community hospital in Beijing was collected. It comprises of 225 records containing 14 features (Table 1) and 775 records containing 4 features. Patients with abnormal results in first step will be required to do further checking. In this study we selected the 14 features 225 records as our investigation objects. The data contains long-term tracking records for 150 healthy individuals and 25 patients with liver disease.

Table 1
Liver function test items.

Item	Normal range
Alanine aminotransferase (ALT)	0–40
Aspartate aminotransferase (AST)	0–40
AST/ALT	0.80–1.5
Gamma glutamyl transpeptidase (GGP)	7–32
Alkaline phosphotase (ALP)	53–128
Total bilirubin (TBILI)	5.1–19.0
Direct bilirubin (DBILI)	0.1–5.1
Indirect bilirubin (IBILI)	5.0–12.0
Total proteins (TP)	60–82
Albumin (ALB)	35–55
Globulin (GLB)	15.0–35.0
ALB/GLB	1.00–2.00
Cholesterol (CHOL)	3.35–6.45
Triglycerides (TRIG)	0.48–1.17

Following the assistance of medical experts, we reconstructed the dataset (Table 2) that includes 175 samples with 13 features and one class label; there are 130 male and 45 female cases; there are 150 normal and 25 abnormal samples. Compared with Table 1, the dataset omits CHOL and TRIG but adds Age and Gender as criteria.

2.2. Visualisation based on dimensionality reduction

PCA [19] is based on the assumption that important signals contribute most to variance and are captured by the first few principal components with a relatively high contribution ratio. PCA is a linear coordinate transformation that rotates a coordinate system in a particular way that does not suffice to extract all of the important signals. ICA [20] comprises a broad field of techniques that aim to decorrelate and maximise the independence of the data with linear mixing of a suitable underlying matrix whose rows are mutually independent and not normally distributed.

Two novel nonlinear methods have been proposed to tackle the dimensionality reduction problem, namely LLE [21] and Isomap [22]. Both of these methods attempt to preserve the local neighbourhood of each object. An unsupervised learning algorithm, LLE, is based on the intuition that even though the columns are points in n-dimensional Euclidean space, the points might actually lie on a much lower dimensional manifold. Isomap [22] determines which points are neighbours on the manifold, finds the intrinsic geometry of the data (as captured in the geodesic manifold

Table 2
Liver function dataset.

Feature	Description
Age	Integer
Gender	Male: 1; female: 0
ALT	Integer
AST	Integer
AST/ALT	Real
GGP	Integer
ALP	Integer
TBILI	Real
DBILI	Real
TP	Real
ALB	Real
GLB	Real
ALB/GLB	Real
Label	Normal: 1 abnormal: 0