COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE 113 (2014) 465–473

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

A random forest classifier for lymph diseases

Ahmad Taher Azar[a,*], *Hanaa Ismail Elshazly*[b,c], *Aboul Ella Hassanien*[b,c], *Abeer Mohamed Elkorany*[b]

[a] *Faculty of Computers and Information, Benha University, Egypt*
[b] *Faculty of Computers and Information, Cairo University, Egypt*
[c] *Scientific Research Group in Egypt (SRGE), Egypt*

ARTICLE INFO

Article history:
Received 28 June 2013
Received in revised form
3 November 2013
Accepted 6 November 2013

Keywords:
Machine learning (ML)
Feature selection (FS)
Genetic algorithm (GA)
Random forest classifier (RFC)
Lymph diseases

ABSTRACT

Machine learning-based classification techniques provide support for the decision-making process in many areas of health care, including diagnosis, prognosis, screening, etc. Feature selection (FS) is expected to improve classification performance, particularly in situations characterized by the high data dimensionality problem caused by relatively few training examples compared to a large number of measured features. In this paper, a random forest classifier (RFC) approach is proposed to diagnose lymph diseases. Focusing on feature selection, the first stage of the proposed system aims at constructing diverse feature selection algorithms such as genetic algorithm (GA), Principal Component Analysis (PCA), Relief-F, Fisher, Sequential Forward Floating Search (SFFS) and the Sequential Backward Floating Search (SBFS) for reducing the dimension of lymph diseases dataset. Switching from feature selection to model construction, in the second stage, the obtained feature subsets are fed into the RFC for efficient classification. It was observed that GA-RFC achieved the highest classification accuracy of 92.2%. The dimension of input feature space is reduced from eighteen to six features by using GA.

1. Introduction

Computer aided diagnosis (CAD) systems have been used for many years. Medical diagnosis is known to be subjective and depends not only on the available data but also on the experience of the physician. It has been suggested that computer translation may hold part of the solution for processing the physician's interpretation [1,2]. Machine learning techniques are increasingly introduced to construct the CAD systems owing to its strong capability of extracting complex relationships in the biomedical data [3,4]. The high volume of medical data requires some helpful classification approaches to support the analysis of this data. Accuracy of classification algorithms used in disease diagnosing is certainly an important issue to be considered. Most medical data has the characteristic of high dimensionality datasets [5,6]. High dimensional data, in general, requires the extraction of most descriptive or discriminative features to be selected and hence the dimension of dataset is reduced [7]. In this context, dimension reduction plays an important role in diagnosing systems to remove irrelevant features from a data set [8,9]. Dimension reduction procedure is useful to decrease dataset complexity with the possible advantage of increased classification performance. Removing the number of irrelevant features for model implementation makes screening tests faster, more convenient and less costly. The current research work is focused on the determination of an optimal feature subset for

* *Corresponding author.*
E-mail addresses: ahmad.azar@fci.bu.edu.eg, ahmad_t_azar@ieee.org (A.T. Azar), hanosoma3002@yahoo.com (H.I. Elshazly), aboitcairo@gmail.com (A.E. Hassanien), a.korani@fci-cu.edu.eg (A.M. Elkorany).
0169-2607/$ – see front matter © 2013 Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.cmpb.2013.11.004

lymphography dataset in order to improve the diagnosis accuracy. The choice is a trade-off between computational time and quality of the generated feature subset solutions.

The lymphatic system aids the immune system in removing and destroying waste, debris, dead blood cells, pathogens, toxins, and cancer cells. It absorbs fats and fat-soluble vitamins from the digestive system and delivers these nutrients to the cells of the body where they are used by the cells. Also, it removes excess fluid, and waste products from the interstitial spaces between the cells. The lymphatic system consists of thin-walled lymphatic vessels, lymph nodes, and two collecting ducts [10]. Lymph vessels are closely associated with the circulatory system vessels. Larger lymph vessels are similar to veins. Lymph capillaries are scatted throughout the body. Contraction of skeletal muscle causes movement of the lymph fluid through valves. Lymph nodes are round or kidney-shaped, and range in size from very tiny to 1 in. in diameter. They are usually found in groups in different places throughout the body, including the neck, armpit, chest, abdomen, pelvis, and groin. Lymph nodes are garrisons of B, T and other immune cells. About two thirds of all lymph nodes and lymphatic tissue are within or near the gastrointestinal tract. The role of these nodes to filter the lymph before it can be returned to the circulatory system. Although these nodes can increase or decrease in size throughout life, any nodes that has been damaged or destroyed, does not regenerate.

The state of the lymphatic system can be detected by lymphography medical imaging techniques [11]. Magnetic resonance lymphography holds much promise for the noninvasive evaluation of lymph nodes. The technique utilizes ultrasmall superparamagnetic particles of iron oxide and has been shown to be highly sensitive and specific in the diagnosis of malignant lymph nodes [12]. The current state of lymph nodes with extracted data from lymphography technique can ascertain the classification of the investigated finding [13]. The enlargement of lymph nodes can be an index to trivial conditions and extends to more significant conditions that threats life [14]. Additionally the status of the lymph nodes could also suggest the occurrence of cancer [9]. Therefore, the main contribution of this paper is to investigate the effectiveness of RFC in conducting the lymph disease diagnostic problem. Aiming at improving the efficiency and effectiveness of the classification accuracy for lymph disease diagnosis, a CAD system based on RFC is introduced. The difference between this study and other studies that address the same topic is that a strong classifier system has been created by combining GA feature selection and random forest decision tree methods, which has very important implications for dimension reduction and sound classification to discriminate between normal and abnormal cases. Furthermore, this method yields more efficient results than any of the other methods tested in this paper.

The structure of the paper is the following: Introduction and related research are briefly described in Sections 1 and 2. Section 3 explains theoretical approach of feature selection methods and RFC. The evaluation procedure is described in Section 4. The dataset and the experimental results are presented in Section 5. Finally, Conclusion and future directions are summarized in Section 6.

2. Related work

Detection of Lymph disease is a prevalent research topic in the literature. Polat and Gunes [15] proposed a novel hybrid classification system based on C4.5 decision tree classifier and one-against-all approach to classify the multi-class problems including dermatology, image segmentation, and lymphography datasets taken from UCI (University of California Irvine) machine learning database. In this work, C4.5 decision tree was initially executed for all the classes of datasets and they reported 84.48%, 88.79%, and 80.11% classification accuracy for dermatology, image segmentation, and lymphography datasets using 10-fold cross validation, respectively. The proposed method based on C4.5 decision tree classifier and one-against-all approach obtained 96.71%, 95.18%, and 87.95% for the above datasets, respectively. Iannello et al. [16] proposed decomposition methods named as One-per-Class (OpC), Error-Correcting Output Codes (ECOC), PairWise Coupling (PWC) for multiclass classification including lymph disease dataset. The comparison has been carried out by employing three different paradigms for the basic classifiers like Multi-Layer Perceptron (MLP) as a neural network, a Nearest Neighbor (NN) as a statistical classifier, and a Support Vector Machine (SVM) as a kernel machine. The experimental results for MLP achieved 82.90%, 79.32% and 75.84% using OpC, ECOC and PWC, respectively. For NN classifier, the results achieved 79.22%, 76.81% and 76.99% using OpC, ECOC and PWC, respectively. Finally, the performance results obtained by SVM using OpC, ECOC and PWC were 87.85%, 81.36%, and 79.44%, respectively. Some of the recent classification results obtained by other studies for Lymph disease dataset are presented in Table 1.

3. Genetic algorithm (GA) and random forest classifier: preliminaries

This section provides a brief explanation of the basic framework of genetic algorithm and random forest classifier, along with some of the key definitions.

3.1. *Genetic algorithm (GA)*

GA is a stochastic search method for solving optimal solutions within large and complicated search spaces. It's a popular type of evolutionary algorithm (EA) that has been successfully used for feature selection. The technique is based on ideas from Darwin's theory of natural selection and "survival of the fittest" [27]. Genetic algorithm operates on a set of individuals called population, where each individual is an encoding of the problem's input data and are called chromosomes. Each chromosome is composed of genes, each of them has a binary value that indicates the presence or not of a specific element of the set. The search for the best solution is guided by an objective function called fitness function. The selected solutions of higher fitness function are more ability to produce new solutions than the less of fitness value while those of weak fitness function will be eliminated gradually. Fitness function controls the selection of best solution and provides a criteria