Available online at www.sciencedirect.com
ScienceDirect
Information Sciences 177 (2007) 771–786
INFORMATION SCIENCES
AN INTERNATIONAL JOURNAL
www.elsevier.com/locate/ins

Thalassaemia classification by neural networks and genetic programming

Waranyu Wongseree [a], Nachol Chaiyaratana [a,*], Kanjana Vichittumaros [b], Pranee Winichagoon [b], Suthat Fucharoen [b]

[a] *Research and Development Centre for Intelligent Systems, Department of Electrical Engineering, Faculty of Engineering, King Mongkut's Institute of Technology North Bangkok, 1518 Piboolsongkram Road, Bangsue, Bangkok 10800, Thailand*
[b] *Thalassaemia Research Centre, Institute of Science and Technology for Research and Development, Mahidol University, Nakhonpathom 73170, Thailand*

Received 23 May 2005; received in revised form 21 March 2006; accepted 3 July 2006

Abstract

This paper presents the use of a neural network and a decision tree, which is evolved by genetic programming (GP), in thalassaemia classification. The aim is to differentiate between thalassaemic patients, persons with thalassaemia trait and normal subjects by inspecting characteristics of red blood cells, reticulocytes and platelets. A structured representation on genetic algorithms for non-linear function fitting or STROGANOFF is the chosen architecture for genetic programming implementation. For comparison, multilayer perceptrons are explored in classification via a neural network. The classification results indicate that the performance of the GP-based decision tree is approximately equal to that of the multilayer perceptron with one hidden layer. But the multilayer perceptron with two hidden layers, which is proven to have the most suitable architecture among networks with different number of hidden layers, outperforms the GP-based decision tree. Nonetheless, the structure of the decision tree reveals that some input features have no effects on the classification performance. The results confirm that the classification accuracy of the multilayer perceptron with two hidden layers can still be maintained after the removal of the redundant input features. Detailed analysis of the classification errors of the multilayer perceptron with two hidden layers, in which a reduced feature set is used as the network input, is also included. The analysis reveals that the classification ambiguity and misclassification among persons with minor thalassaemia trait and normal subjects is the main cause of classification errors. These results suggest that a combination of a multilayer perceptron with a blood cell analysis may give rise to a guideline/hint for further investigation of thalassaemia classification.

Keywords: Thalassaemia classification; Neural network; Genetic programming

* Corresponding author. Tel.: +66 2 913 2500x8410; fax: +66 2 585 6149.
E-mail addresses: nchl@kmitnb.ac.th, nachol1@go.com (N. Chaiyaratana).

0020-0255/$ - see front matter
doi:10.1016/j.ins.2006.07.009

1. Introduction

Thalassaemia is a genetic disease that causes a reduction in the life span of a red blood cell [28]. The disease is a result of an abnormality in the genes that regulate the formation of *haemoglobin* (Hb)—a core component of the red blood cell. In order to make the diagnosis, the blood characteristics must be analysed. A *complete blood count* (CBC) is the primary screening test for a laboratory diagnosis of thalassaemia. However, there is still a limitation in the analysis of data due to a large number of possible candidate characteristics. In addition, there are various types of thalassaemia and thalassaemia trait. (Persons with thalassaemia trait do not have the disease but inherit genes that cause the disease.) As a result, a manual diagnostic process can only be carried out by specialists whose decision is based upon an index from mathematically combined values of blood characteristics [14,5].

Early attempts to formulate an automated diagnostic tool employed image analysis [18], statistical [6] and clustering techniques [3]. Later, the implementation protocol has shifted to the expert systems, in which both rule-based [23,24,17] and hybrid neural network/rule-based systems [4] have been successfully tested in clinical trials. Nonetheless, these tools broadly differentiate between a wide range of blood-related diseases including various types of anaemia. In order to narrow the diagnostic target down to the differentiation between thalassaemic patients, persons with thalassaemia trait and normal subjects, an alternative automated diagnostic tool is required. Recently, a successful implementation of a neural network [1,2], a k-nearest neighbour technique [2] and a support vector machine [2] as a thalassaemic diagnostic tool has been reported. However, the tool can only differentiate between two types of thalassaemic gene carriers and normal subjects. Further works are required in order to expand the tool capability to cover all major types of thalassaemic patients and persons with thalassaemia trait that are commonly found in Thailand where the total number of types is much larger [7].

The thalassaemia classification can generally be formulated into a pattern recognition problem. The input patterns or samples in this case would be blood-related data covering the characteristics of red blood cells, *reticulocytes* (young red blood cells that usually remain in the bone marrow with only a few venturing out into the circulating blood) and platelets. These characteristics can be directly obtained from an automatic blood cell analyser. On the other hand, the target classification output would be either the disease/trait type or a normal-subject flag. The use of a neural network and a genetic programming (GP) based decision tree as the classifiers is proposed. GP techniques have been successfully used as an evolvable classifier [15] and an optimisation tool for evolving neural networks [27] and fuzzy systems [9,26,27] in classification tasks. However, GP-constructed classifiers have rarely been used as medical decision support tools unlike neural networks which have been successfully used in numerous medical data classification tasks [8,10].

This paper is organised as follows. The thalassaemia classification problem of interest is explained in Section 2. In Section 3, the description of the neural network and GP-based classifier is given. Next, the classification results are discussed in Section 4. Finally, the conclusions are drawn in Section 5.

2. Thalassaemia classification problem

In this section, the basic background about thalassaemia and the data set used in the classification task will be explained; the background will be covered in Section 2.1 while the details about the data set will be given in Section 2.2.

2.1. Background on thalassaemia

Thalassaemia is a form of chronic anaemia that reduces the life span of red blood cells [28]. The disease stems from an abnormality in the genes that regulate the formation of a protein called *globin*, which is a major component of haemoglobin. Each red blood cell contains approximately 300 million molecules of haemoglobin. Hence, a change in the structure of globin affects the structure and functionality of a red blood cell. A globin molecule contains two parts: *α-globin* and *β-globin*. The α-globin contains 141 amino acids, which are regulated by genes on chromosome 16. The β-globin consists of 146 amino acids, which are governed by genes on chromosome 11. Since the regulatory genes reside on two *autosomes*, the transmission mode of