Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 69 (2015) 132 – 142

Procedia Computer Science

7th International Conference on Advances in Information Technology

Comparison of Classifiers for the Risk of Diabetes Prediction

Nongyao Nai-arun[a,*], Rungruttikarn Moungmai[a]

[a]*Faculty of Science and Technology, Nakhon Sawan Rajabhat University, Thailand.*

Abstract

This paper applied a use of algorithms to classify the risk of diabetes mellitus. Four well known classification models that are Decision Tree, Artificial Neural Networks, Logistic Regression and Naive Bayes were first examined. Then, Bagging and Boosting techniques were investigated for improving the robustness of such models. Additionally, Random Forest was not ignored to evaluate in the study. Findings suggest that the best performance of disease risk classification is Random Forest algorithm. Therefore, its model was used to create a web application for predicting a class of the diabetes risk.

Peer-review under responsibility of the Organizing Committee of IAIT2015

Keywords: diabetes; random forest; logistic regression; artificial neural networks; decision tree; naive bayes; bagging; boosting.

1. Introduction

Diabetes mellitus (DM) is a chronic non-communicable disease. The disease has been closely followed by World Health Organization (WHO) and International Diabetes Federation (IDF) since worldwide number of diabetes increase continuously. It was found that there were 387 million people with diabetes in year 2014 and have a tendency to be 592 million patients in the next 20 years[1]. IDF also found that almost half of diabetes in South East Asia is undiagnosed. According to these amounts, the disease should be controlled and properly maintained for efficient and sustainable prevention.

The annual report 2013 of Department of Disease Control, Ministry of Public Health, reports that diabetes is the top three of chronic non-communicable diseases in Thailand[2,3]. The statistics shows that 1 of 13 adults Thais had

* Corresponding author. Tel.: +6-681-643-8016; fax: +6-656-882-531
E-mail address: nongyao25@hotmail.com, r.moungmai@gmail.com

Peer-review under responsibility of the Organizing Committee of IAIT2015
doi:10.1016/j.procs.2015.10.014

diabetes and the total number of people with diabetes is not less than 3 million. In the future, there will be more than 7 million people are at risk of diabetes[4]. The report also indicates that the number of diabetes is likely to be increased every year. Consider diabetes death rate, there are about 12 dead with diabetes in every 100 thousand people. This can be seen that the rate is a small number however this amount is only the dead with diabetes. In fact, diabetes is an important cause of other diseases such as stroke and heart diseases which are the top three of chronic non-communicable diseases and have high death rates[3]. It also leads to the destruction of cells in the body such as nerves, blood vessels, heart, eyes and kidneys[4,5].

Nowadays, the situation of diabetes in Thailand has been concerned due to 1 of 3 diabetes patients is undiagnosed and unaware. It was also found that the age of patients trends to decrease. Moreover, the number of female is more likely than male with diabetes and the patients are obese people more than the non-obese[5].

As mentioned above, a study of disease classification is considered since it holds great potential for improving human health and personal treatment. In this paper, four popular classifiers for disease risk prediction are studied. These algorithms consists Decision Tree, Artificial Neural Network, Logistic Regression and Naïve Bayes. After that Bagging and Boosting techniques are combined with those algorithms to improve the robustness of each model. At the end, Random Forest algorithm is applied.

The objective of this study is to predict the risk of diabetes for everyone without the need of blood test or going to a hospital. The study also aims to encourage and promote good health of people. In addition, the diabetes prediction will be created as a simple diagnosis application and will be published by a website. However, this application is only an initial diagnosis. People who found that they are in the diabetes risk group should go to see a doctor for formal diagnosis to prevent themselves from serious diabetes.

2. Material and Method

The format of this study is as shown in Fig. 1.

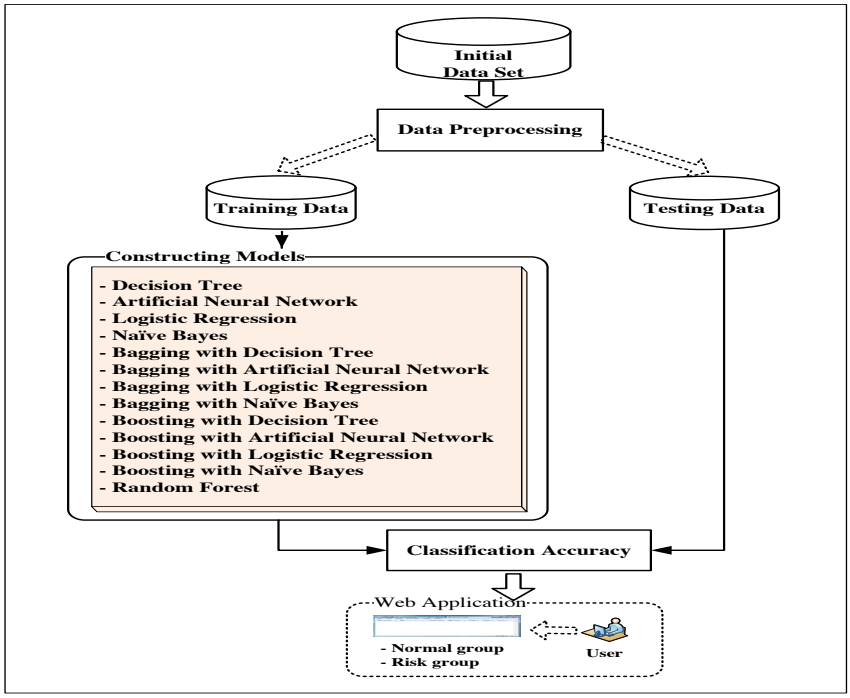

Fig. 1. The process of conceptual framework.