Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 82 (2016) 115 – 121

Procedia Computer Science

Symposium on Data Mining Applications, SDMA2016, 30 March 2016, Riyadh, Saudi Arabia

Performance Analysis of Data Mining Classification Techniques to Predict Diabetes

Sajida Perveen[a]*, Muhammad Shahbaz[a], Aziz Guergachi[b], Karim Keshavjee[c]

[a]*Department of Computer Science & Engineering, University of Engineering & Technology, Lahore, Pakistan*
[b]*Ted Rogers School of Information Technology Management, Ryerson University, Toronto, Ontario, Canada*
[c]*University of Victoria, School of Health Informatics, Victoria, British Columbia, Canada*

Abstract

Diabetes Mellitus is one of the major health challenges all over the world. The prevalence of diabetes is increasing at a fast pace, deteriorating human, economic and social fabric. Prevention and prediction of diabetes mellitus is increasingly gaining interest in healthcare community. Although several clinical decision support systems have been proposed that incorporate several data mining techniques for diabetes prediction and course of progression. These conventional systems are typically based either just on a single classifier or a plain combination thereof. Recently extensive endeavors are being made for improving the accuracy of such systems using ensemble classifiers. This study follows the adaboost and bagging ensemble techniques using J48 (c4.5) decision tree as a base learner along with standalone data mining technique J48 to classify patients with diabetes mellitus using diabetes risk factors. This classification is done across three different ordinal adults groups in Canadian Primary Care Sentinel Surveillance network. Experimental result shows that, overall performance of adaboost ensemble method is better than bagging as well as standalone J48 decision tree.

Peer-review under responsibility of the Organizing Committee of SDMA2016
Keywords: Diabetes Mellitus; Ensemble method; Base Learner; Bagging; Adaboost and Decision tree

* Corresponding authors. Tel.: +92-556601721.
E-mail address: sajida.uaar@gmail.com

Peer-review under responsibility of the Organizing Committee of SDMA2016
doi:10.1016/j.procs.2016.04.016

1. Introduction

Diabetes mellitus (DM), commonly known as diabetes, is a chronic and one of the dramatically increasing metabolic diseases in the world[6, 11]. It is associated with an abnormal increase in the level of glucose (hyperglycemia) in blood, ensued either owing to the inadequate production of insulin by pancreas (Type 1 diabetes) or the cells failure in effective response to insulin produced by pancreas (Type 2 diabetes)[13]. The downside of all this variability in plasma glucose (hyperglycaemia, hypoglycemia) is that it leads to severe damage to many of the body's vital systems especially blood vessels and the nervous system[10]. While its causes are not yet entirely understood, scientists believe that both genetic factors and environmental triggers are involved therein[8]. However, diabetes used to be most prevalent in adults and once called "adult-onset" diabetes. It is now widely believed that diabetes mellitus is closely related with the aging process.

According to Canadian Diabetes Association (CDA), between 2010 and 2020, the number of people diagnose with diabetes in Canada is expected to escalate from 2.5 million to about 3.7 million[7]. Unfortunately, worldwide the picture is no different from this. According to the International Diabetes Federation, number of individuals with diabetes mellitus has reached 382 million in 2013[14] that bring 6.6% of the world's total adult population with diabetes. Health care expenditures for diabetes are anticipated to be $490 billion for 2030, accounting for 11.6% of the total health care expenditures in the world[2]. Furthermore, diabetes is a potentially independent contributing risk factor to microvascular complications. Its patients are likely to be more vulnerable to an elevated risk of microvascular damage thereby exposing them to cardio vascular disease two to fourfold more as compared to no diabetic individuals. This micro vascular damage and consequent cardio vascular disease ultimately lead to retinopathy, nephropathy and neuropathy[8]. Studies revealed that the life expectancy for people with diabetes might get curtailed by as much as 15 years[17].

Given the above narrated consequences, early stage detection and diagnosis of diabetes is the need of the day. In this context, Electronic Medical Records (EMRs) play a crucial role by keeping track of repeated clinical measurements related to particular patient's condition over time. To provide a rapid and minutely detailed analysis of medical data, diabetes risk scoring models as well as their various algorithms has been widely investigated. Schwarz et al. [16] provided a comprehensive survey of these models with their specificity and sensitivity. However, as these risk scoring models involve human intervention though to some extent in deciding criteria and risk score, it may expose the results to the human error.

Data mining is a prominent tool set in medical databases. This promising approach improves sensitivity and/or specificity of disease detection and diagnosis by opening a window of comparatively better resources. It also substantially reduces accompanied cost by bypassing unwanted and expensive medical tests[9]. Extensive studies regarding diabetes prediction has been undergone for several years. Recently, some reports have compared different learning techniques. Such comparisons are generally a few and conducted on Pima Indian diabetic database with a limited number of data sets.

On the other hand, this study follows the adaboost and bagging Data Mining ensemble techniques using J48 (c4.5) decision tree as a base learner along with standalone data mining technique J48 (c4.5). More specifically, the dataset used in this study for disease diagnosis and decision making is obtained from the Canadian Primary Care Sentinel Surveillance Network (CPCSSN) database. That is Canada's first multi-disease EMR-based surveillance system. Firstly; The objective of this study is to evaluate the performance of aforementioned techniques of data mining to accurately classify patients with diabetes mellitus using diabetes risk factors across three different ordinal adults groups in CPCSSN, namely (i) young adults (ii) middle aged adults (iii) adults older than 55. Secondly; to identify the best ensemble framework for J48 decision tree that would help identify the diabetes patients efficiently and most importantly, with high accuracy. The rest of paper is organized as follows: Section 2 presents material and method. Section 3 describes results, evaluation and discussion. Conclusion is given in section 4.