Chapter 17
A Method for Classification Using Data Mining Technique for Diabetes: A Study of Health Care Information System

Ahmad Al-Khasawneh
Hashemite University, Jordan

ABSTRACT

Many researchers in the health information system field have been attracted to develop computer applications that help in the diagnosis process. Imperatively, data mining algorithms address the vital role in all of these applications. Many contributions were made in this area. There has always been a debate on the algorithm that gives the best classifier, the parameters to be used, the dataset pre-processing steps, etc. In this paper, the author largely emphasizes that the best way to build a predictive model with relatively high classification accuracy is to build several predictive models and to choose the model that gives the best results through parameters optimization. Diagnosing diabetes mellitus has gained considerable attention in the last few decades due to the increased severity of the disease. In this research, the author reviews four predictive data mining approaches that are being used in diagnosing diabetes. Four models were implemented to diagnose diabetes from PIMA dataset; k-nearest neighbour, support vector machine, multilayer perceptron neural network, and naive bayesian network. Giving the highest classification accuracy, support vector machine technique outperformed the others with a value of 78.83%.

1. INTRODUCTION

Diabetes is a chronic disease that results when the percentage of sugar in blood exceeds its normal levels. This is the case when sugar is not absorbed well by body cells. This could be due to the inability of the pancreas to produce enough insulin (type1) or the inability of the body cells to respond to the produced insulin (type2) (IDF Diabetes Atlas, 2013). As the number of diabetes cases has increases remarkably over the last few decades, many researchers have been attached to develop software systems that help clinicians do their job more professionally especially in the diagnosis process.

DOI: 10.4018/978-1-5225-0788-8.ch017

In health care, data mining plays a vital role in the medical applications including diagnosis, prognosis, and therapy. Applying data mining in health care applications is usually referred to as clinical data mining (CDM) (Jacob & Ramani, 2012). Clinical data mining involves the conceptualization, extraction, analysis, and interpretation of the available clinical data for practical knowledge-building, clinical decision making, and partition reflection (Jacob & Ramani, 2012).

Among the various medical applications, data mining mainly targets the diagnosis ones (Al-Khasawneh & Hijazi 2014). To diagnose a disease is to decide whether a patient suffers from a specific disorder depending on the medical signs, symptoms, and tests. Computer programs used to help in this aid are called clinical decision support systems (CDSSs), or more specifically diagnosing decision support systems (DDSSs).

A medical diagnosis is a classification problem (Saidi, Chikh, & Settouti, 2011). Hence, the majority of the CDSS employs predictive data mining to diagnose a disease (Al-Khasawneh & Hijazi 2014). Predictive data mining is a supervised model building algorithm (Williams, 2011) which tries to predict trends and future behaviours depending on historical variables (Omari, 2013) and values wherein the probable values of the outcome are specified previously. The goal of predictive data mining in the diagnosis process is to build models from old observations or historical data (i.e. usually patients' records) to predict the outcome of new patients or observations to help in the clinical decision making process. In the predictive data mining, the data set consists of instances, each instance is characterized by attributes or features and another special attribute represents the outcome variable or the class (Bellazzi & Zupanb, 2008).

Often, the goal of any data mining project is to build a model from the available data. Thus, data mining models are objective models rather than subjective since it is driven by the available data. Predictive data mining builds both classification and regression modelling using several algorithms such as decision trees, random forests, boosting, support vector machines, linear regression, and neural networks (Williams, 2011) & (Al-Khasawneh & Hijazi 2014). Descriptive data mining uses cluster analysis and association rules modelling techniques (Williams, 2011).

Indeed, the majority of data mining projects (including diagnosis) are predictive and employs predictive modelling techniques. Classification models predict the class of a new observation among predefined categories of the target variable (Williams, 2011), whilst the output of the regression modelling is a numeric value rather than a class (Williams, 2011).

To diagnose diabetes, we need to classify diabetic form non-diabetic patients. In this paper, we introduce several predictive modelling approaches that could help in this classification. Four models have been implemented to diagnose diabetes; k-nearest neighbour, support vector machine, multilayer perceptron neural network, and naive bayesian network. All of the models were implemented from the Pima Indian diabetes dataset and validated using 10-cross validation techniques.

The paper is structured as follows; section 2 summarizes the works in the literature that are most relevant to this work. Section 3 introduces the proposed approach including preparing the dataset, the implemented models, and the performance analysis. Lastly, the paper is concluded in section 4.

2. RELATED WORK

Applying data mining techniques in solving clinical problems has attracted many researchers especially in the diagnosis process. In this section, we introduce significant works wherein predictive modelling are utilized to help in disease diagnosis process, specifically, the diabetes illness.