Chapter 11
Coronary Heart Disease Prognosis Using Machine-Learning Techniques on Patients With Type 2 Diabetes Mellitus

Angela Pimentel
FCT-UNL, Portugal

Hugo Gamboa
FCT-UNL, Portugal

Isa Maria Almeida
APDP-ERC, Portugal

Pedro Matos
APDP-ERC, Portugal

Rogério T. Ribeiro
APDP-ERC, Portugal

João Raposo
APDP-ERC, Portugal

ABSTRACT

Heart diseases and stroke are the number one cause of death and disability among people with type 2 diabetes (T2D). Clinicians and health authorities for many years have expressed interest in identifying individuals at increased risk of coronary heart disease (CHD). Our main objective is to develop a prognostic workflow of CHD in T2D patients using a Holter dataset. This workflow development will be based on machine learning techniques by testing a variety of classifiers and subsequent selection of the best performing system. It will also assess the impact of feature selection and bootstrapping techniques over these systems. Among a variety of classifiers such as Naive Bayes (NB), Random Forest (RF), Support Vector Machine (SVM), Alternating Decision Tree (ADT), Random Tree (RT) and K-Nearest Neighbour (KNN), the best performing classifier is NB. We achieved an area under receiver operating characteristics curve (AUC) of 68,06% and 74,33% for a prognosis of 3 and 4 years, respectively.

DOI: 10.4018/978-1-5225-7122-3.ch011

1. INTRODUCTION

About 387 million people worldwide have diabetes. In 2014 diabetes was the direct cause of 4.9 million deaths (WHO, 2015). The prevalence of diabetes is increasing in developed and developing countries and is predicted to achieve 7,7% worldwide by 2030 (Shaw, 2010). Within Europe, Portugal is reported as having the highest prevalence of diabetes. Over 1 million Portuguese have diabetes, with almost half of these still undiagnosed (Gardete-Correia, 2010). Diabetes increases the risk of heart disease and stroke. In a multinational study, 50% of people with type 2 diabetes (T2D) die of cardiovascular disease (primarily heart disease and stroke) (Morrish, 2001). The cardiovascular events associated with T2D and the high incidence of other macrovascular complications, such as peripheral vessel disease and amputations, are a major burden of disease and a huge economic determinant. Clinicians and health authorities for many years have expressed interest in identifying individuals at increased risk of CHD. The underlying cause of CHD is a slow buildup of plaques after fatty deposits on the inner wall of the blood vessels that supply the heart muscle with blood (the coronary arteries). These fatty deposits gradually obstruct the arteries, sometimes clogging them and subsequently reduce the flow of blood to the heart. There is no single cause for CHD, but there are risk factors that increase the chance of developing it. A useful clinical diagnostic tool for following cardiac diseases evolution is with the electrocardiogram (ECG)-Holter records. Long term Holter monitoring is used for patients with heart conditions such as arrhythmias. Heart beats with unusual timing or unusual ECG morphology can be very helpful in early diagnosis of hearts with damaged electrophysiology.

In clinical training and practice, prognosis typically receives less attention than diagnosis and disease treatment. Yet many clinical decisions are not fully useful unless the patient's prognosis is considered (Gill, 2012) (Lan, 2012).

The healthcare industry has generated large amounts of data, driven by record keeping, compliance and regulatory requirements, and patient care (Raghupathi, 2016). With the potential to improve the quality of healthcare delivery, while reducing costs, these massive quantities of data (known as 'big data') supports a wide range of medical and healthcare functions, which includes clinical decision support, disease surveillance, and population health management. (Burghard, 2012). Machine learning enables the extraction of implicit, previous unknown and potentially useful information from data (Witten, 2011). With machine learning techniques, supervised or unsupervised methods are applied in order to extract and evaluate data patterns which can be used to take better decisions and to present the knowledge we extracted in a better way.

This paper proposes a new prognostic approach of CHD for T2D patients based on a Holter dataset. This approach, based on machine learning methods, supports the applicability of using the Holter dataset, to efficiently predict T2D patients that are likely to develop CHD. We also highlight the importance of using feature selection algorithms for the construction of our predictive model. A drawback of this study is the low number of patients that evolve to CHD, therefore the use of boostrapping technique. The result is a prognostic tool that allows the identification of high-risk subjects in the diabetic population. The entire data set captures 8 years (2006-2014) of clinical visits, and a CHD prognosis can be made with intervals for time windows of 3 and 4 years, with AUC values of 68,06% and 74,33%, respectively using a Naïve Bayes classifier. We also show that our predictive model outperforms the Framingham study (D'Agostino, 2000) for this population. We present a solution with a higher number of features than the