Feature Selection and Classification Model Construction on Type 2 Diabetic Patient’s Data

Yue Huang¹, Paul McCullagh¹, Norman Black¹, and Roy Harper²

¹ School of Computing and Mathematics, Faculty of Engineering, University of Ulster, Jordanstown, BT37 0QB Northern Ireland, UK
ythuang@infj.ulst.ac.uk, {PJ.McCullagh, ND.Black}@ulst.ac.uk
² The Ulster Hospital, Dundonald, Belfast, BT16 0RH, Northern Ireland, UK

Abstract. Diabetes is a disorder of the metabolism where the amount of glucose in the blood is too high because the body cannot produce or properly use insulin. In order to achieve more effective diabetes clinic management, data mining techniques have been applied to a patient database. In an attempt to improve the efficiency of data mining algorithms, a feature selection technique ReliefF is used with the data, which can rank the important attributes affecting Type 2 diabetes control. After selecting suitable attributes, classification techniques are applied to the data to predict how well the patients are controlling their condition. Preliminary results have been confirmed by the clinician and this provides optimism that data mining can be used to generate prediction models.

1 Introduction

Diabetes is a major global health problem, affecting around 194 million people worldwide, and that number is expected to increase to 300 million by 2025. In attempt to investigate a patient’s prognosis, data mining techniques have been applied to a ‘working’ Type 2 diabetic database to generate information that can be verified by the clinician and possibly provide new knowledge. The data source is a secondary care database system [1]. Specifically, this research concentrates on:

1. Identifying significant factors influencing diabetes control;
2. Predicting individuals in the population with poor diabetes control status based on physiological and examination factors.

1.1 Data Mining

The role of data mining is to extract interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from large amounts of data. It comprises a variety of techniques used to identify nuggets of information or decision-making knowledge in data, and extracting these in such a way that they can be put to use in areas such as decision support, prediction and estimation. The data is often voluminous but, as it stands, is of low value as no direct use can be made of it; it is the...
hidden information in the data that is useful. Data mining includes pre-processing, applying a data-mining algorithm, and post-processing the results.

1.2 Diabetes and HbA1c

According to a World Health Organisation, on a global scale, there are around 194 million people with diabetes in the adult population, 50% of which are undiagnosed [2]. Diabetes occurs when the body cannot produce enough insulin or the insulin produced is ineffective. There are two main types of diabetes: Type 1 and Type 2. The most common form is Type 2 diabetes, previously called non-insulin-dependent diabetes mellitus (NIDDM). Since its morbidity term can be up to 30 years depending on disease control, it greatly affects patient’s quality of life and the prevention of its complications is very important. It has been shown that better blood glucose control will reduce the risk of complications significantly [3-6].

HbA1c is a laboratory test, which reveals average blood glucose over the previous 12 weeks. This test is recommended by the American Diabetes Association [7] to monitor long term glucose control. HbA1c is usually recorded every 3 months, but may be performed more often, if needed. Specifically, it measures the number of glucose molecules attached to haemoglobin, a substance in red blood cells. According to the investigation of UK Prospective Diabetes Study group (UKPDS), every 1% HbA1c reduction means 35% less complication risk, both for the micro- and macrovascular complications. Microvascular complications increase dramatically when the HbA1c measurement is over 10% [8].

1.3 Data Mining in Diabetes

Due to the greatly increased amount of data gathered in medical databases traditional manual analysis has become inadequate, and methods for efficient computer-based analysis are indispensable. To address this problem, knowledge discovery in databases (KDD) methods are being developed to identify patterns within the data that can be exploited. Data mining methods have been applied to a variety of medical domains in order to improve medical decision making; diagnostic and prognostic problems in oncology, liver pathology, neuropsychology, and gynaecology. Improved medical diagnosis and prognosis may be achieved through automatic analysis of patient data stored in medical records, i.e., by learning from past experiences [9].

Data analysis of diabetes has previously been reported. Breault et al.[10] examined a diabetic data warehouse using the Classification and Regression Tree (CART) system to deduce a series of association rules among different attributes. They emphasise that bad glycaemia control is primarily associated with a younger age, rather than the related complications or whether patients have related diseases. Their work shows that data mining can discover novel associations that are useful to clinicians and administrators. Although the classification accuracy achieved by CART was just 59.5%, the authors found it useful to derive rule sets.

Miyaki et al. [11] also adopted the CART algorithm to identify the factors that influence diabetic vascular complications. They used AnswerTree [12] to build the