Chapter 14
A Fast Feature Selection Method Based on Coefficient of Variation for Diabetics Prediction Using Machine Learning

Tengyue Li
University of Macau, Macau

Simon Fong
University of Macau, Macau SAR

ABSTRACT

Diabetes has become a prevalent metabolic disease nowadays, affecting patients of all age groups and large populations around the world. Early detection would facilitate early treatment that helps the prognosis. In the literature of computational intelligence and medical care communities, different techniques have been proposed in predicting diabetes based on the historical records of related symptoms. The researchers share a common goal of improving the accuracy of a diabetes prediction model. In addition to the model induction algorithms, feature selection is a significant approach in retaining only the relevant attributes for the sake of building a quality prediction model later. In this article, a novel and simple feature selection criterion called Coefficient of Variation (CV) is proposed as a filter-based feature selection scheme. By following the CV method, attributes that have a data dispersion too low are disqualified from the model construction process. Thereby the attributes which are factors leading to poor model accuracy are discarded. The computation of CV is simple, hence enabling an efficient feature selection process. Computer simulation experiments by using the Prima Indian diabetes dataset is used to compare the performance of CV with other traditional feature selection methods. Superior results by CV are observed.

DOI: 10.4018/978-1-6684-6291-1.ch014

INTRODUCTION

Diabetes is a global health concern in both developed and developing countries, and its prevalence is rising. In just UK alone, 2.9 million people are suffering from diabetes mellitus in 2011 that constitutes to 4.45% of the population (Holman et al., 2011). By 2025, it is projected to have 5 million people in UK inflicted with diabetes. This incurable metabolic disorder is chronic and characterized by deficiency of insulin secretion or insensitivity of the body tissues to insulin. The former is known as Type-I insulin-dependent diabetes mellitus (IDDM) where the body defects to produce sufficient insulin due to autoimmune destruction of pancreatic β-cells. As a result, the patients' body cells may wither because they cannot absorb the needful amount of glucose in the bloodstream without this important hormone. The second type is called Type-II non-insulin-dependent diabetes mellitus which is usually associated with obesity and lack of bodily exercises. It will inevitably lead to insulin treatment, probably for life-long. Early detection of diabetes has become vital and the detection techniques are maturing over the years. However, it is reported that about half of the patients with Type II diabetes are undiagnosed and the latency from disease onset to diagnosis may exceed over a decade (American diabetes association) (International Diabetes Federation). Therefore, the importance of early prediction and detection of diabetes that enables timely treatment of hyperglycaemia and related metabolic abnormalities is escalating.

In the light of this motivation, diabetes prediction models are being formulated and developed in machine-learning research community that claimed to be able to do blood glucose prediction based on the historical records of diabetes patients and their relevant attributes. One of the most significant works is by Jan Maciejovski (Maciejowski, 2002) who formulated predictive diabetic control by using a group of linear and non-linear programming functions that take into consideration of variables and constraints. The other direction related to blood glucose prediction is time-series forecasting (Ståhl & Johansson, 2009), which take into account of the measurements of the past blood glucose cycles, in order to do some short-term blood glucose forecasts. Another popular choice of algorithm in implementing a blood glucose predictor is artificial neural network (Otto et al., 2000; Gogou et al., 2001; Akmal et al., 2011) which non-linearly maps daily regimens of food, insulin and exercise expenditure as inputs to a predicted output. Although neural network predictors usually can achieve a relatively high accuracy (88.8% as in (Akmal et al., 2011)), the model itself is a black-box where the logics in the process of decision making are mathematical inference. For example, numeric weights associated in each neuron and the non-linear activation function. Recently some researchers advocated applicability of decision trees in predicting diabetic prognosis such as batch-training model (Han et al., 2009) and real-time incremental training model (Zhang et al., 2012). The resultant decision tree is in a form of predicate logics IF-THEN-ELSE rules which are descriptive enough for decision support when the rules are embedded in some predictor system, as well as for reference and studies by clinicians. However, one major drawback on decision tree is the selection of the appropriate data attributes or features that should be general enough to model the historical cases, while providing sufficiently high prediction accuracy in the event of unseen case.

Potentially there exist many factors (so-called features) for analysis and diagnosis of the diabetes of patients; these factors may be direct physiological symptoms or lifestyle habits that contribute to the disease. However, there is no standard rule-of-thumb in deciding which of these factors into the inclusion of the model induction (Janecek et al., 2008), given different physicians might have their own opinions. At convenience when all the available features are included in the process of model construction, quite often some of these features may found to be insignificant or irrelevant. Consequently, the accuracy of the prediction model reduces because these the inappropriate feature might have added randomness to