Informatics in Medicine Unlocked 2 (2016) 92–104

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu

Rule extraction using Recursive-Rule extraction algorithm with J48graft combined with sampling selection techniques for the diagnosis of type 2 diabetes mellitus in the Pima Indian dataset

Yoichi Hayashi*, Shonosuke Yukita

Department of Computer Science, Meiji University, Tama-ku, Kawasaki, Kanagawa 214-8571, Japan

ARTICLE INFO

Article history:
Received 19 December 2015
Received in revised form
29 January 2016
Accepted 24 February 2016
Available online 23 April 2016

Keywords:
Rule extraction
Type 2 diabetes mellitus
Re-RX algorithm
Sampling selection
Pima Indian diabetes
Data mining

ABSTRACT

Diabetes is a complex disease that is increasing in prevalence around the world. Type 2 diabetes mellitus (T2DM) accounts for about 90–95% of all diagnosed adult cases of diabetes. Most present diagnostic methods for T2DM are black-box models, which are unable to provide the reasons underlying diagnosis to physicians; therefore, algorithms that can provide further insight are needed. Rule extraction can provide such explanations; however, in the medical setting, extracted rules must be not only highly accurate, but also simple and easy to understand. The Recursive-Rule eXtraction (Re-RX) algorithm is a "white-box" model that provides highly accurate classification. However, due to its recursive nature, it tends to generate more rules than other algorithms. Therefore, in this study, we propose the use of a rule extraction algorithm, Re-RX with J48graft, combined with sampling selection techniques (sampling Re-RX with J48graft) to achieve highly accurate, concise, and interpretable classification rules for the Pima Indian Diabetes (PID) dataset, which comprises 768 samples with two classes (diabetes or non-diabetes) and eight continuous attributes. The use of this algorithm resulted in an average accuracy of 83.83% after 10 runs of 10-fold cross validation. Sampling Re-RX with J48 graft achieved substantially better accuracy and provided a considerably fewer average number of rules and antecedents than the original Re-RX algorithm. These results suggest that sampling Re-RX with J48graft provides more accurate, concise, and interpretable extracted rules than previous algorithms, and is therefore more suitable for medical decision making, including the diagnosis of T2DM.

1. Introduction

Diabetes is a complex disease characterized by a lack of or resistance to insulin, a hormone critical for the regulation of blood sugar. In healthy individuals, the pancreas produces insulin to help metabolize sugar in the blood and keep blood glucose (sugar) levels within a normal range. Diabetics cannot produce or are resistant to insulin, and as a result, are unable to remove glucose from their bloodstreams. Consequently, glucose levels in the blood increase, leading to serious health problems [1].

In 2011, there were 347 million diabetics worldwide, and by 2030, this number is expected to increase to 552 million. About 4.6 million deaths were caused by diabetes in 2011, and by 2030, it is projected to be the seventh leading cause of death [2].

According to the Centers for Disease Control and Prevention, an estimated 29.1 million people, or 9.3% of the US population, have diabetes [3], 8.1 million of whom remain undiagnosed. In 2010, diabetes was listed as the underlying cause of death on 69,071 death certificates and a cause of death another 234,051, making it the seventh leading cause of death in the US.

Diabetes can affect the entire body and is associated with severe complications such as heart disease, stroke, vision loss, kidney failure, and lower-limb amputations. Good glucose control can help avoid some complications, particularly microvascular eye, kidney, and nerve disease, and early detection and treatment can help prevent disease progression; therefore, monitoring that includes dilated eye exams, urine tests, and foot exams is essential. Because diabetics and prediabetics are at an increased risk of cardiovascular disease, blood pressure and lipid management, and especially smoking cessation, are particularly important.

There are two main clinical classifications of diabetes: type 1 and type 2. Onset of type 1 diabetes, which was previously known as insulin-dependent diabetes mellitus or juvenile-onset

* Corresponding author.
E-mail addresses: hayashiy@cs.meiji.ac.jp (Y. Hayashi), redgtvo9606@gmail.com (S. Yukita).

http://dx.doi.org/10.1016/j.imu.2016.02.001

diabetes, accounts for about 5% of all diagnosed adult cases of diabetes. Although it can occur at any age, the peak age for diagnosis of type 1 diabetes is in the mid-teens.

The peak age of onset of type 2 diabetes mellitus (T2DM), which was previously known as non–insulin-dependent diabetes mellitus or adult-onset diabetes, is typically later than that of type 1 diabetes and accounts for about 90–95% of all diagnosed adult cases of diabetes. T2DM usually starts with insulin resistance, a disorder in which cells primarily within the muscles, liver, and fat tissue do not utilize insulin properly. The beta cells in the pancreas begin to gradually lose the ability to produce sufficient quantities of insulin as the need for the hormone increases. In contrast to beta cell dysfunction, the role of insulin resistance differs among individuals; some primarily have insulin resistance and only a minor defect in insulin secretion, while others primarily have a lack of insulin secretion and only slight insulin resistance.

Although the exact causes of complex diseases such as T2DM have yet to be identified [4], a combination of genetic, environmental, and lifestyle factors is suspected [5]. An ever-increasing amount of data is being collected in medical databases, and historical data on complex diseases, such as patients' blood glucose levels, is becoming more widely available; therefore, traditional methods of manual analysis have become inadequate. As a result, a variety of data mining techniques are being applied in order to discover new patterns of disease and promote the early detection and diagnosis of complex diseases such as diabetes [6].

The diagnosis of T2DM is a two-class classification problem, and numerous methods for diagnosing T2DM have been successfully applied to the classification of different tissues. However, most present diagnostic methods [1,7–47] for T2DM are black-box models. A drawback of black-box models is that they cannot adequately reveal information that may be hidden in the data.

For example, even in cases for which high-performance classifiers [2,4,8,24,25,32,33] allow the accurate assignment of instances to groups, black-box models are unable to provide the reasons underlying that assignment to physicians; therefore, algorithms that can provide insight into these underlying reasons are needed. Rule extraction can provide such explanations, and is it therefore becoming increasingly popular. However, in the medical setting, extracted rules must be not only highly accurate, but also simple and easy to understand. Rules are one of the most popular symbolic representations of knowledge discovered from data, and are more comprehensible, particularly "black boxes" like unseen medical datasets, than other representations [48].

The Recursive-Rule eXtraction (Re-RX) algorithm, originally intended to be a rule extraction tool, was recently developed by Setiono et al. [49]. Re-RX provides a hierarchical, recursive consideration of discrete variables prior to analysis of continuous data, and can generate classification rules from neural networks (NNs) that have been trained on the basis of both discrete and continuous attributes.

In contrast to black-box models, the Re-RX algorithm [49] is a "white-box" model that provides highly accurate classification. It is easy to explain and interpret in accordance with the concise extracted rules associated with IF-THEN forms. Due to its ease of understanding, the Re-RX algorithm is typically preferred by both physicians and clinicians alike.

However, due to its recursive nature, the Re-RX algorithm tends to generate more rules than other rule extraction algorithms. Therefore, one of the major drawbacks of the Re-RX algorithm is that it typically generates expansive extraction rules for middle-sized or larger datasets.

It is important to consider both accuracy and interpretability for extracted classification rules. The number of correctly classified test samples typically determines the accuracy of each extracted classification rule, while the number of extracted rules and the average number of antecedents in the extracted rules determines their interpretability.

To achieve both concise and highly accurate extracted rules while maintaining the good framework of the Re-RX algorithm, we recently proposed supplementing the Re-RX algorithm with J48graft [51], a class for generating a grafted C4.5 decision tree [50]. J48graft [52] is the result of the C4.5A [53] algorithm being implemented in open source data mining software referred to as the "all-tests-but-one partition (ATBOP)" [53]. In Re-RX with J48graft, J48graft [52] is employed to form decision trees in a recursive manner, while multi-layer perceptrons (MLPs) are trained using backpropagation (BP), which allows pruning [54], thereby generating more efficient MLPs for highly accurate rule extraction. Re-RX with J48graft provides rules that are not only highly accurate, but also easily explained and interpreted in terms of the concise extracted rules; that is, Re-RX with J48graft provides IF-THEN rules. This white-box model is easier to understand and is therefore often preferred in the medical setting.

In this study, we first proposed the use of a rule extraction algorithm, Re-RX with J48graft [51], combined with sampling selection techniques (sampling Re-RX with J48graft) [55,56] for preprocessing. We then investigated the accuracy, conciseness, and interpretability of diagnostic rules extracted for the Pima Indian Diabetes (PID) dataset using sampling Re-RX with J48graft based on a comparison with both crisp rule extraction techniques [21,27,28] and previous fuzzy rule extraction techniques [1,12–16,29–31,43]. As a typical example of T2DM, we used the PID dataset from the repository of machine learning at the University of California Irvine (UCI) [57]. The PID dataset comprises 768 samples with two classes (diabetes or non-diabetes) and eight continuous attributes. Important values missing from the PID dataset are discussed in Section 3.7.

In Section 5, we review the performance of rule extraction algorithms for the PID dataset since 2003, and compare the previous extracted fuzzy and crisp rules with the performance of the present extracted rules. In Sections 5.1–5.6, we compare the concrete rules for the PID dataset extracted by the proposed algorithm with those obtained using the four kinds of previous rule extraction algorithms recommended by the American Diabetes Association (ADA) for the diagnosis of diabetes. In Section 5.7, we also compare the classification accuracy obtained by the proposed algorithm with that obtained by other classifier systems for the PID dataset.

We explain the role of the oral glucose tolerance test (OGTT) and body mass index (BMI) for the diagnosis of the PID dataset in Section 6.1, and discuss the interpretation of rules extracted by the proposed algorithm from the perspective of medical informatics in Section 6.2. In Section 6.3, we discuss the trade-offs between accuracy and the number of extracted rules using trade-off curves, and in Section 6.4, we elucidate the trade-offs between accuracy and the average number of antecedents. Finally, we provide a summary and conclusion in Section 7.

2. Related works

In 1996, Shanker [10] evaluated the effectiveness of artificial NN (ANN) classifiers in predicting the onset of non–insulin-dependent diabetes mellitus among the Pima Indian female population. According to Knowler et al., the Pima Indians have the highest reported incidence of diabetes in the world [58]. Smith et al. [59] used the same dataset to test a model for predicting the onset of diabetes mellitus. In this study, ANNs were used to model the relationship between the onset of diabetes mellitus and various risk factors for diabetes among Pima Indian women.