Informatics in Medicine Unlocked 1 (2015) 1–8

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

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

Use of the recursive-rule extraction algorithm with continuous attributes to improve diagnostic accuracy in thyroid disease

Yoichi Hayashi *, Satoshi Nakano, Shota Fujisawa

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

ARTICLE INFO

Article history:
Received 17 October 2015
Received in revised form
29 December 2015
Accepted 29 December 2015
Available online 15 February 2016

Keywords:
Thyroid disease diagnosis
Re-RX algorithm
Rule extraction
Decision tree

ABSTRACT

Thyroid diseases, which often lead to thyroid dysfunction involving either hypo- or hyperthyroidism, affect hundreds of millions of people worldwide, many of whom remain undiagnosed; however, diagnosis is difficult because symptoms are similar to those seen in a number of other conditions. The objective of this study was to assess the effectiveness of the Recursive-Rule Extraction (Re-RX) algorithm with continuous attributes (Continuous Re-RX) in extracting highly accurate, concise, and interpretable classification rules for the diagnosis of thyroid disease. We used the 7200-sample Thyroid dataset from the University of California Irvine Machine Learning Repository, a large and highly imbalanced dataset that comprises both discrete and continuous attributes. We trained the dataset using Continuous Re-RX, and after obtaining the maximum training and test accuracies, the number of extracted rules, and the average number of antecedents, we compared the results with those of other extraction methods. Our results suggested that Continuous Re-RX not only achieved the highest accuracy for diagnosing thyroid disease compared with the other methods, but also provided simple, concise, and interpretable rules. Based on these results, we believe that the use of Continuous Re-RX in machine learning may assist healthcare professionals in the diagnosis of thyroid disease.

1. Introduction

About 200 million people worldwide, or nearly 15% of the entire adult population, are affected by thyroid diseases. In the United States, thyroid diseases affect about 27 million people, half of whom remain undiagnosed. Thyroid diseases often lead to thyroid dysfunction involving either hypo- or hyperthyroidism, which are both relatively prevalent among the general population [1]. Hypothyroidism, a condition in which the thyroid gland is underactive and stops producing adequate levels of thyroid hormone, is more prevalent, accounting for about 80% of diagnosed cases. The other 20% are commonly diagnosed as hyperthyroidism, a condition in which the thyroid gland is overactive and produces excessive levels of thyroid hormone. Every major organ in the body is affected by thyroid function; therefore, disorders of the thyroid gland are a matter of great importance.

Two active hormones produced by the thyroid gland, triiodothyronine (T_3) and levothyroxine (LT_4), play a wide range of important roles in the body, including protein production and the regulation of body temperature. In order to help cells convert oxygen and nutrients into energy, T_3 and LT_4 production must be within normal ranges, which are typically based on their concentrations in blood. In this study, we calculated the normal ranges for T_3, LT_4, thyroid stimulating hormone (TSH), thyroxine (T_4) utilization rate, and free thyroxine index (FTI).

Thyroid diseases can be difficult to diagnose because the associated symptoms are similar to those seen in a number of other conditions. However, thyroid disorders can be identified using a TSH test, even before symptom onset [2].

Diagnosing thyroid disease is a three-class classification problem, and numerous supervised methods for diagnosing thyroid disease have been successfully applied to the classification of different types of thyroid dysfunction [1–20].

However, many current diagnostic methods [1–13,15,18,19] for thyroid disease 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 the case that a method allows the accurate assignment of instances to groups, it is unable to provide the reasoning underlying that assignment to users. Systems and/or algorithms that can provide insight into these underlying reasons are needed. Among the supervised methods, rule extraction is capable of providing such explanations, and is therefore becoming increasingly popular. However, it is necessary, particularly in the medical setting, that extracted rules are not only simple and easy to understand, but also highly accurate.

* Corresponding author. Tel.: +81 44 934 7475; fax: +81 44 931 5161.
E-mail addresses: hayashiy@cs.meiji.ac.jp (Y. Hayashi), me.sa.nakano@gmail.com (S. Nakano), hoiminn627@gmail.com (S. Fujisawa).

http://dx.doi.org/10.1016/j.imu.2015.12.003

The number of extracted classification rules and the average number of antecedents determines their interpretability, while the number of correctly classified test samples typically determines the accuracy.

The objective of this study was to develop an improved rule extraction algorithm for a large and highly imbalanced medical dataset, i.e., the Thyroid dataset. In machine learning and data mining, learning classification rules from examples is one of the oldest and most common tasks. Such rules are typically expressed as symbolic descriptions in an "IF (conditions) THEN (target class)" form, in which conditions are created as a conjunction of elementary tests on values of attributes that describe learning examples, and the assignment of an example satisfying the condition to a given class is indicated by the rule consequence. Rules are one of the most popular symbolic expressions of knowledge derived from data, and they have been described as being more comprehensible and interpretable than other representations, particularly black-box models such as medical datasets [21].

In the present study, we proposed using the Re-RX algorithm [22] with continuous attributes (Continuous Re-RX) and attempted to extract highly accurate, concise, and interpretable classification rules for the diagnosis of thyroid disease. The Thyroid dataset was obtained from the University of California Irvine (UCI) Machine Learning Repository [23] and comprises both discrete and continuous, i.e., mixed, attributes. Two versions of the Thyroid dataset have been used for benchmarking in previous studies. One version comprises 7200 samples [10,14–17], while the other comprises 215 samples [2,3,5,6,18–20].

For the purposes of this study, we used the 7200-sample Thyroid dataset. Continuous Re-RX is capable of handling such mixed-attribute datasets. Therefore, the objective of this study was to assess the accuracy and comprehensibility of rules for the Thyroid dataset using Continuous Re-RX based on a comparison with both types of classification rule sets extracted by Duch et al. [14].

2. Related research

Numerous supervised methods have been developed for the diagnosis of thyroid disease, including the following: extreme learning machines [1]; support vector machines [2–4,20]; neural networks (NNs) [5–8,14,15]. decision trees [5], k-nearest neighbor classifiers [9], fuzzy classifiers [16,17]; hybrid case-based reasoning [18]; mixture of expert models [10]; immune algorithms [11], immune recognition systems [12]; neuro-fuzzy expert systems [13], and differential evolution [19].

Some researchers have experimented with extracting Boolean rules from NNs [24–26], which has led to encouraging results that exhibit good performance, a reduced number of rules, relevant input variables, and increased interpretability. However, these methods use Boolean rules, and therefore do not extract continuous rules.

Setiono et al. [22] proposed a Recursive-Rule Extraction (Re-RX) algorithm for rule extraction from an NN trained for solving a classification problem having mixed discrete and continuous input data attributes. This algorithm shares some similarities with other existing rule extraction algorithms.

In the Re-RX algorithm [22], the C4.5 decision tree [27] is frequently employed in a recursive manner, while multilayer perceptron ensembles (MLPs) are trained using backpropagation NNs; this allows pruning [28] and therefore generates more efficient MLPs for highly accurate rule extraction.

The Re-RX algorithm 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 physicians and clinicians alike.

Results regarding the extraction of classification rules for diagnosing thyroid disease have been reported in a number of studies [14,16,17,20]. Among these studies, Duch et al. [14] reported highly accurate classification for the Thyroid dataset and provided two types of relatively simple and concrete classification rule sets.

3. Theory

3.1. Recursive-rule extraction algorithm: Re-RX algorithm

The Re-RX algorithm [22] is designed to generate classification rules from datasets that have both discrete and continuous attributes. The algorithm is recursive in nature and generates hierarchical rules. The rule conditions for discrete attributes are disjointed from those for continuous attributes. The continuous attributes only appear under the conditions of the rules that are lowest in the hierarchy. The outline of the algorithm is as follows:

Re-RX Algorithm (S, D, C)
Input: A set of data samples S having discrete attributes D and continuous attributes C.
Output: A set of classification rules.
1. Train and prune [28] a NN using the dataset S and all of its D and C attributes.
2. Let D' and C' be the sets of discrete and continuous attributes, respectively, still present in the network, and let S' be the set of data samples correctly classified by the pruned network.
3. If $D'=\phi$, then generate hyperplane to split the samples in S' according to the values of the continuous attributes C', and then stop.

Otherwise, use only the discrete attributes D' to generate the set of classification rules R for dataset S'.

4. For each rule, R_i is generated:

If support $(R_i) > \delta_1$ and error $(R_i) > \delta_2$, then

- Let S_i be the set of data samples that satisfies the condition of rule R_i, and let D_i be the set of discrete attributes that does not appear in rule condition R_i.
- $D_i=\phi$, then generate hyperplane to split the samples in Si according to the values of their continuous attributes C_i, and then stop.
- Otherwise, call Re-RX (S_i, D_i, C_i).

The support of a rule is the percentage of samples that are covered by that rule. The support and the corresponding error rate of each rule are checked in Step 4. If the error exceeds the threshold appropriate δ_2 and the support meets the maximum appropriate threshold δ_1, then the subspace of this rule is further subdivided either by recursively calling Re-RX when discrete attributes are still absent in the conditions of the rule or by generating a separating hyperplane that involves only the continuous attributes of the data.

3.2. Mechanism of the Re-RX algorithm

To allow a better understanding of the mechanism underlying the Re-RX algorithm, we provide a brief overview and explore the concept behind its design, which has not been previously described in detail, in Fig. 1. We used C4.5 [26] to generate decision trees in the Re-RX algorithm. An overview of the Re-RX algorithm is shown in Fig. 1.