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.

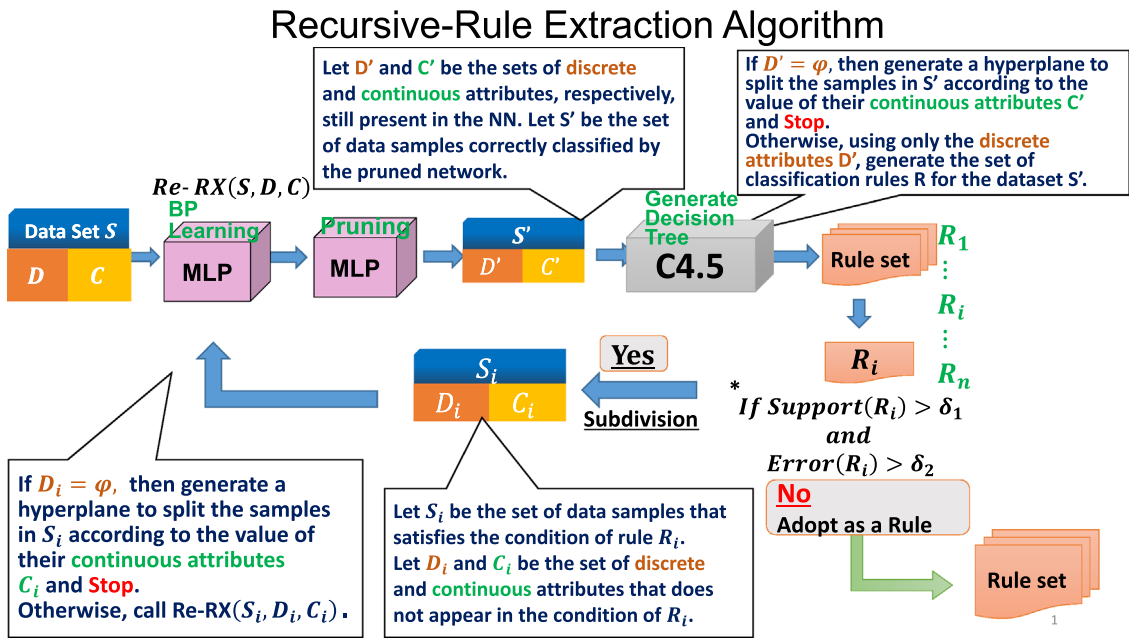

Fig. 1. Schematic of the Recursive-Rule Extraction (Re-RX) algorithm MLP: multi-layer perceptron; NN: neural network; BP: back-propagation.

In Fig. 1, the subdivision of the Re-RX algorithm is a unique function and inherent in its nature. Each subsequent subdivision allows the use of other unused attributes, which increases both the number and accuracy of extracted rules.

Needless to say, accuracy, comprehensibility, and conciseness in extracted rules have important trade-offs. Extracted rules before subdivision are more concise and interpretable, yet have lower accuracy, whereas extracted rules after subdivision are less concise, but have better accuracy.

A major advantage of the Re-RX algorithm recently developed by Setiono et al. [22] is that it was intended as a rule extraction tool and provides a hierarchical, recursive consideration of discrete variables prior to analysis of continuous data. In addition, it is capable of generating classification rules from NNs that have been trained on the basis of discrete and continuous attributes.

In other words, the Re-RX algorithm achieves highly accurate rule extraction and offers good comprehensibility through the generation of perfect or strict separation between discrete and continuous attributes in the antecedent of each extracted rule.

3.3. C-MLP2LN and separability split value (SSV)

To facilitate extraction of logical rules from an MLP network, it should be smoothly transformed into a logical network (LN), which is a network that performs logical operations. This transformation, known as MLP2LN [29], may be realized in several ways. If the goal is to find logical rules for a previously trained large MLP network, skeletonization is the preferred method. Otherwise, starting from a single neuron and directly constructing the LN using training data (the C-MLP2LN [29] method) is faster and more accurate. However, interpretation of the activation of the MLP network nodes is difficult; therefore, a smooth transition from the MLP to a type of LN performing similar functions is advocated.

The split value (or cutoff point) is defined differently for continuous and discrete features. In the case of continuous features, the split value is a real number, while in other cases, it is a subset of the set of alternative values of the feature.

The higher the separability of a split value, the better. Points beyond the borders of feature values existing in the dataset have an SSV (separability split value) [29] equal to zero, while the separability of all points between the borders is positive. This means that every dataset containing vectors that belong to at least two different classes, a maximal SSV exists for each feature that has at least two different values. When the feature being examined is continuous and several maximal SSVs are close to each other, the split value closest to the mean tends to be selected. To avoid such situations, SSVs that are natural for a given dataset, i.e., values that are between adjacent feature values, are examined. If there are two maxima with smaller SSVs in between, or if the feature is discrete, then the selection of the best SSV may be arbitrary. If context-independent linguistic variables are desired, the separability criterion can be used in several different ways to discretize a continuous feature.

4. Materials and methods

4.1. Re-RX algorithm with continuous attributes (Continuous Re-RX)

A primary goal of the Re-RX algorithm is the strict separation of discrete and continuous attributes in extracted rules; however, we found that this design often leads to reduced accuracy. As shown in Fig. 2, the Re-RX algorithm eliminates the continuous attributes (C) before the C4.5 decision tree is generated. In the present paper, we propose using both discrete (D) and continuous attributes (C) to generate the decision tree [30]. Although this seems to be counterintuitive with the design concept of the Re-RX algorithm, in that it results in the generation of a more complex decision tree, the use of both types of attributes is done to enhance accuracy. An outline of Continuous Re-RX is as follows:

Continuous Re-RX (S, D, C)
Input: A set of data samples (S) having both discrete (D) and continuous (C) attributes.
Output: A set of classification rules.
1. Train and prune [28] an NN using the dataset S and all of its D and C attributes.

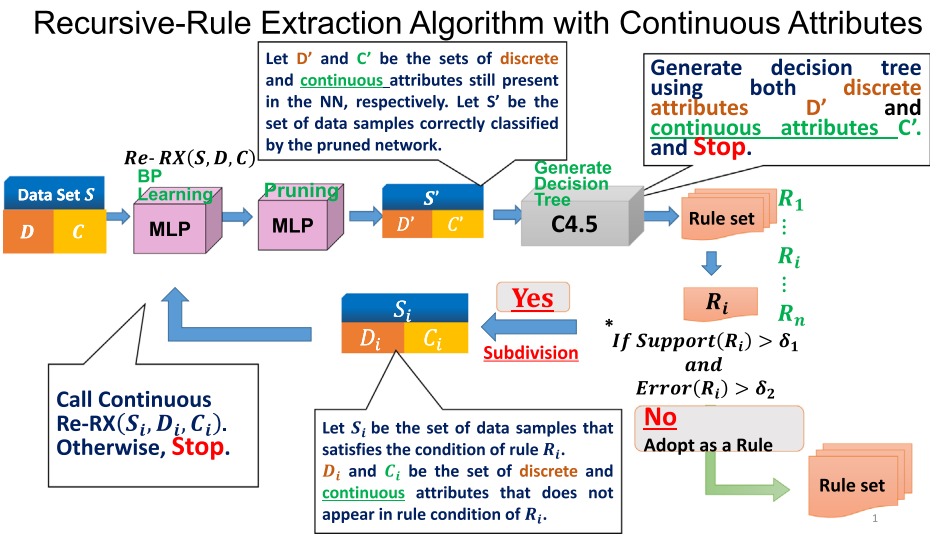

Fig. 2. Schematic of the Recursive-Rule Extraction (Re-RX) algorithm with continuous attributes (Continuous Re-RX) MLP: multi-layer perceptron; NN: neural network; BP: back-propagation.

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. Generate decision tree by using both discrete (D') and continuous (C') attributes [30].
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, let D_i be the set of discrete attributes, and let C_i be the set of continuous attributes that does not appear in rule condition R_i.
- Call Continuous Re-RX (S_i, D_i, C_i).

Otherwise, Stop.

As with the Re-RX algorithm, the support of a rule in the proposed Continuous Re-RX 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.

As shown in Fig. 2, to avoid such difficulties in Continuous Re-RX, we carefully set the value of subdivision rate and the values of δ_1 and δ_2 in Step 4. However, because this cannot be done with other algorithms, 10-fold cross validation (CV) is unable to be conducted for the Thyroid dataset due to the small numbers of samples. To the best of our knowledge, 10-Fold CV results from the Thyroid dataset have not been previously reported.

4.2. Thyroid dataset and experimental setup

The Thyroid dataset is a big dataset composed of screening tests for thyroid symptoms. The training and test data in this set comprise 3772 and 3428 medical records, respectively, in which the thyroid is classified as normally functioning (Class 3), underfunctioning (primary hypothyroidism, Class 2), or overactive (hyperthyroidism, Class 1). Hyper- and hypothyroidism represent 2.3% (166 cases) and 5.1% (367 cases) of the dataset, respectively. The remaining 92.6% (6667 cases) are classified as normal.

The dataset is highly imbalanced, and thus it is difficult to classify thyroid levels. A total of 21 attributes with 15 binary and six continuous variables were used to determine the appropriate classification for each of the 7200 cases. Each record contains information on the patient's age, binary indicators such as gender, history of goiter, current medications, history of pregnancy, history of thyroid surgery and tumor, and levels of the following five hormones: TSH; T_3; LT_4; T_4; and FTI [31,32].

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, Duch et al. [14] reported highly accurate classification for the Thyroid dataset and provided two types of relatively simple and concrete classification rule sets.

If a dataset is characterized by an unequal distribution between classes, it is considered to be imbalanced (known as between-class imbalance), and no consensus has been reached regarding the degree of imbalance between class cardinalities. Some studies have focused on data containing one class several times smaller than other classes, while other studies have investigated more severe imbalance ratios (e.g., 1:10, 1:100, or even greater). The lack of data (absolute rarity), i.e., the number of examples in the rare (minority) class, is too small to properly detect regularities in the data, and this is a critical issue in classification and/or rule extraction. The imbalance can be either intrinsic (i.e., it is a direct result of the nature of the data), or it can be the result of a costly acquisition of examples from the minority class. For the purposes of this study, we considered that the difficulty regarding the lack of sufficient learning examples in the minority class did not only apply to binary (two-class) problems, but also to multiclass data with between-class imbalance [21].

A variety of algorithms have been developed to extract classification rules. Although these algorithms have been shown to be effective in solving numerous learning and classification problems, some data characteristics may be problematic and reduce the performance of extracted classifiers. One such problem may be related to class imbalances in the set of learning examples. In imbalanced data, one class (the minority class) contains a much smaller number of examples than the other classes (the majority classes). However, examples from the minority class are usually of primary interest, and thereby their correct recognition is of primary importance. This type of situation frequently occurs in medical diagnosis, where far fewer patients require special attention such as therapy or treatment. The failure to recognize an

illness and assign a proper treatment is often much more dangerous than misdiagnosis, which can be corrected in an additional examination [21].

Numerous solutions [21,33–34] have been proposed to improve the classification performance of classifiers learned from imbalanced data. These are typically divided into the following two general categories: methods operating on the data level, and those operating on the algorithmic level. Focusing on the algorithmic level, Napierala and Stefanowski [21] proposed a new rule induction algorithm called Bottom-up induction of Rules And Cases for Imbalanced Data (BRACID) that aimed to improve learning classifiers from imbalanced data.

As described in Section 4.1, Continuous Re-RX can extract accurate, concise, and interpretable rules from the Thyroid dataset for not only majority classes such as Class 3, but also minority classes such as Classes 1 and 2, using the appropriate values of subdivision rate, δ_1, and δ_2.

5. Results

5.1. Performance

We trained the Thyroid dataset using Continuous Re-RX and obtained the maximum training accuracy (TR ACC) and the maximum test accuracy (TS ACC), the number of extracted rules (# rules), the average number of antecedents (Ave. # ante.), and the area under the receiver operating characteristics curve (AUC) [35] (Table 1). In this paper "ELSE" (the so-called default rule), which appeared in the extracted rule set, was counted as one rule.

Table 1 shows the effectiveness of the subdivision process in Continuous Re-RX. Although Continuous Re-RX before subdivision achieved more accuracy, i.e., 98.51% for the Thyroid dataset, it resulted in a greater number of rules (i.e., five). This is because, generally speaking, there is a trade-off relationship between the accuracy and the number of rules in rule extraction algorithms. Each subdivision in the Re-RX algorithm increased both the accuracy and the number of rules.

The rule set has three rules, but is composed of just two attributes (TSH and T_4). Therefore, after subdivision, as shown in Fig. 3, R_3 became the *new* R_3, R_4, and R_5 by adding discrete attributes (on-thyroxine and Thyroid-surgery). As a result, Continuous Re-RX extracted five rules and achieved 98.51% accuracy for the Thyroid dataset.

The extracted rule set by Continuous Re-RX (before subdivision and after subdivision) is as follows:

The area under the receiver operating characteristics curves (AUC-ROC) for Classes 1 and 2 (0.897), 2 and 3 (0.995), and 1 and 3 (0.893) are shown in Figs. 4–6, respectively (Total AUC=0.928).

5.2. Comparisons

No previous studies have estimated performance results using k-fold cross validation [36] or averaged the results over a series of 10 runs. Therefore, we also compared TR ACC and TS ACC, # rules, and Ave. # ante with three rule extraction and two fuzzy-clustering algorithms (Table 2).

Table 1
Performance of Continuous Re-RX for the thyroid dataset.

Thyroid dataset	TR ACC (%)	TS ACC (%)	# rules	Ave. # ante.	AUC-ROC
Continuous Re-RX before subdivision	97.43	96.70	3	1.67	0.921
Continuous Re-RX after subdivision	99.05	98.51	5	2.8	0.928

Continuous Re-RX: Recursive-Rule Extraction algorithm with continuous attributes; TR: training; TS: Test; ACC: accuracy; Ave.: average; Ante.: antecedents; AUC-ROC: area under the receiver operating characteristics curve.

Before subdivision:

R_1: TSH ≤ 0.006 THEN Class 3

R_2: TSH > 0.006 AND T_4 ≤ 0.054 THEN Class 1

R_3: TSH > 0.006 AND T_4 > 0.054 THEN Class 2

After subdivision:

R_1: TSH ≤ 0.006 THEN Class 3

R_2: TSH > 0.006 AND T_4 ≤ 0.054 THEN Class 1

R_3: TSH > 0.006 AND T_4 > 0.054 AND on-thyroxine = NO AND Thyroid-surgery = NO THEN Class 2

R_4: TSH > 0.006 AND T_4 > 0.054 AND on-thyroxine = NO AND Thyroid-surgery = YES THEN Class 3

R_5: TSH > 0.006 AND T_4 > 0.054 AND on-thyroxine = YES THEN Class 3

Fig. 3. Rule extraction process applied to the Thyroid dataset using Continuous Re-RX and rules extracted in the present study (before subdivision and after subdivision), TSH: thyroid stimulating hormone; T_4: thyroxine.

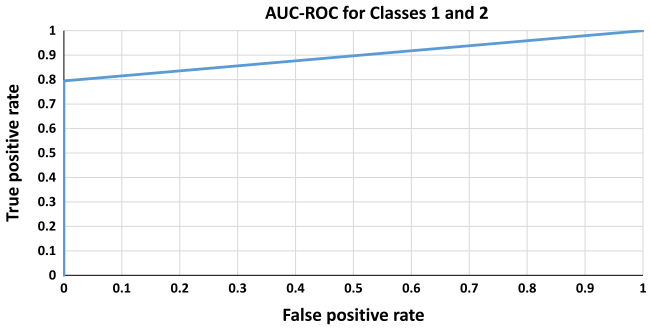

Fig. 4. Area under the receiver operating characteristics curves (AUC-ROC) for Classes 1 and 2.

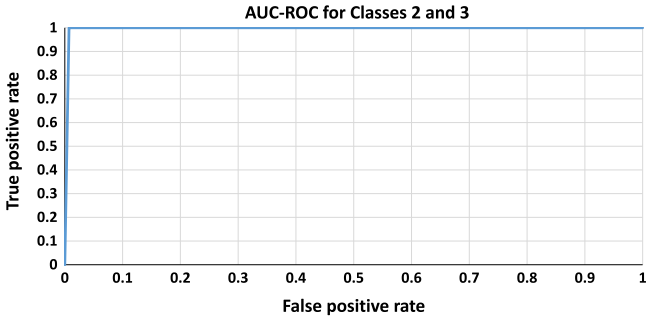

Fig. 5. AUC-ROC for Classes 2 and 3.

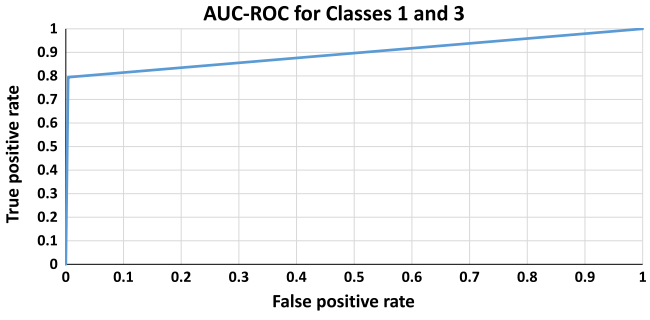

Fig. 6. AUC-ROC for Classes 1 and 3.

Table 2
Comparison of rule extraction algorithms for the Thyroid dataset.

	TR ACC (%)	TS ACC (%)	# of rules	Ave. # ante.
Continuous Re-RX	97.43	96.70	**3**	**1.67**
C-MLP2LN [14]	99.89	99.36	4	2.5
SSV [14]	99.79	99.33	3	2.67
Fuzzy clustering with ellipsoidal regions [16]	98.1	96.79	25	–
Fuzzy clustering with ellipsoidal regions with six continuous attributes [17]	–	95.04–95.22	2–13	–

SSV: separability split value; TR: training; TS: test; ACC: accuracy; Ave.: average; Ante.: antecedent.

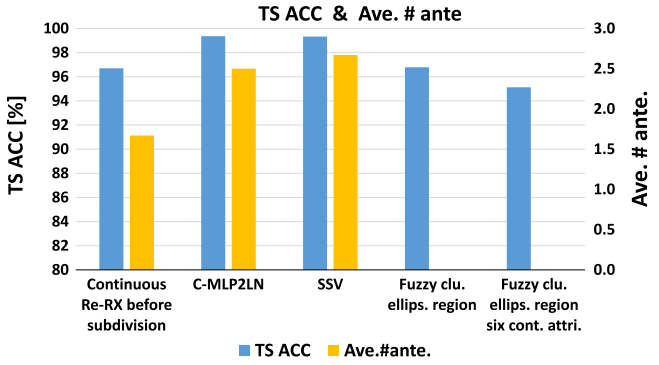

Fig. 7. Bar graph comparing the maximum test accuracy (TS ACC) and the average number of antecedents (Ave. # ante.) obtained using Continuous Re-RX (present method), C-MLP2LN, separability split value (SSV), Fuzzy clustering with ellipsoidal regions (Fuzzy clu. ellips. region) and Fuzzy clustering with ellipsoidal regions with six continuous attributes (Fuzzy clu. ellips. region six cont. attri).

According to Fig. 7, although strictly speaking, it is difficult for the authors to show that the accuracy of the proposed method has no statistical difference ($p > 0.05$) from that of methods reported in previous studies, approximately the same level of accuracy could be seen among Continuous Re-RX, C-MLP2LN and SSV for the Thyroid dataset; however, the number of antecedents is much smaller compared with that of C-MLP2LN and SSV.

Three rules extracted using the SSV reported by Duch et al. [16] showed slightly better classification accuracy (99.33%) than Continuous Re-RX after subdivision. These results suggest that Continuous Re-RX provides approximately the same accuracy as both C-MLP2LN and SSV. The extracted rule sets obtained using Continuous Re-RX, C-MLP2LN, and SSV are shown in Figs. 3, 8, and 9, respectively.

The extracted rule set using C-MLP2LN [14] is as follows:

The extracted rule set using SSV [14] is as follows:

6. Discussion

6.1. Comparisons

Comparison of these four rule sets revealed characteristic class differences between each rule. In the present study, all classes appeared at least once in the extracted rules. The first rule set (before subdivision) shown in Fig. 3 defines Class 3 using only one rule with one attribute (TSH). This means that the TSH value is clearly very important in the diagnosis of Class 3. In the same manner, T_4 is very important in the diagnoses of Classes 1 and 2.

R_{11}: TSH > 0.03048 AND FTI < 0.06427 THEN Class 1

R_{12}: TSH ∈ [0.00602, 0.02953] AND FTI < 0.06427 AND

T_3 < 0.02322 THEN Class 1

R_2: TSH ≥ 0.00602 AND FTI ∈ [0.06427, 0.18671] AND

LT_4 ∈ [0.050, 0.1505] AND on-thyroxine = NO AND

Thyroid-surgery = NO THEN Class 2

ELSE Class 3.

Fig. 8. Rules extracted in a previous study using the C-MLP2LN method TSH: thyroid stimulating hormone; T_3: triiodothyronine; LT_4: levothyroxine; FTI: free thyroxine index.

R_1: TSH > 0.00605 AND FTI < 0.06472 AND

Thyroid-surgery = NO THEN Class 1

R_2: TSH > 0.00605 AND FTI > 0.06472 AND LT_4< 0.1505

AND Thyroid-surgery = NO AND on-thyroxine = NO

THEN Class 2

ELSE Class 3.

Fig. 9. Rules extracted in a previous study using the SSV method TSH: thyroid stimulating hormone; LT_4: levothyroxine; FTI: free thyroxine index.

We believe that the reason three of the five extracted rules define Class 3 is due to the characteristics of the Thyroid dataset, which primarily consists of Class 3 samples.

On the other hand, using C-MLP2LN, Duch et al. [14] reported four rules by which to classify Classes 1 and 2. The samples that do not satisfy the first three rules are classified as Class 3. Focusing on Classes 1 and 2, few records exist in the dataset, and others are classified only as Class 3. The rules extracted by Duch et al. initially appear to successfully classify the records; however, the algorithm they proposed did not extract any classification rules for Class 3, even though these account for the majority of samples (92.6%) in the Thyroid dataset. In contrast, Continuous Re-RX not only achieves the highest classification rule accuracy, but also extracts concrete, concise, and interpretable classification rules for Class 3.

Furthermore, each of the four rules extracted by Duch et al. consisted of 2.5 antecedents, which is more than the 1.67 obtained by Continuous Re-RX before subdivision. Similarly, SSV provided three classification rules with relatively high accuracy (99.33%) [14]. Although this rule set was more concise in terms of number of rules, it had the same problems as C-MLP2LN, in that each extracted rule consisted of 2.67 antecedents, which again, is more than the 1.67 obtained by Continuous Re-RX before subdivision.

Abe and Ruck [16,17] concluded that discrete attributes did not improve classification accuracy when using a fuzzy classifier with ellipsoidal regions for the Thyroid dataset, thereby suggesting the importance of continuous attributes. The results of the present study support their conclusions.

When the characteristics of the Re-RX algorithm are considered, the algorithm appears to give markedly greater priority to discrete than continuous attribute rules; this sacrifices accuracy in order to maintain interpretability. Particularly in cases in which the characteristics of the dataset depend on continuous attributes [16,17], users (physicians) who diagnose using data from the Thyroid dataset do so with higher accuracy, but reduced interpretability.

6.2. Theoretical discussion

We think that increasing the average number of extracted rules is equivalent to adopting the process of subdivision in the Re-RX algorithm. The same property also holds true for Continuous Re-RX.

As described in Section 5.1, when comparing the rule sets before and after subdivision, if a rule set has a higher average number of antecedents, the accuracy is also expected to increase.

However, the higher the number of antecedents, the more complex the extracted rules will be, leading not only to decreased interpretability, but also to a decreased capability of generalization and to overfitting for the test and training datasets, respectively. As a result, the accuracy of the test dataset would be expected to decrease.

Therefore, in Re-RX and Continuous Re-RX, the values of δ_1 and δ_2 can be properly set to avoid increasing the number of rules and antecedents.

6.3. Significance of the present research

Traditionally, symptoms, as well as the degree and burden of suffering associated with both hypo- and hyperthyroidism have lacked clarity and intersubjective validity, and individuals affected by these disorders have faced stigmatization and found it difficult to construct a meaningful identity [37].

It is therefore believed that more accurate screening tools and promotional activities would lead to the proper diagnosis of hypo- or hyperthyroidism for much greater numbers of patients. The diagnostic accuracy of the Thyroid dataset seen in the present study suggests that improving rule extraction methods, even to a slight extent, would help prevent the misdiagnosis of thyroid disease.

Reliable diagnosis guidelines for thyroid dysfunction are currently available as a result of marked advances in the technology for the measurement of T_3, T_4, TSH and the anti-TSH receptor antibody (TRAb), which are important attributes for diagnosis [38–40]. Since our Thyroid dataset does not include TRAb as an input attribute, the three rules extracted using Continuous Re-RX did not include TRAb as an attribute in the antecedents. Our extracted rules for the diagnosis of Thyroid dysfunction are limited to the Thyroid dataset, which was created in the mid-1980s, and therefore remain insufficient.

7. Conclusions

In the present study, rule extraction from a large and highly imbalanced medical dataset using Continuous Re-RX resulted in accurate, concise, and interpretable rules.

The maximum accuracy of the three extracted rules for diagnosing thyroid disease was considerably high (96.70%), and the level of accuracy was approximately the same as that achieved using C-MLP2LN and SSV. Furthermore, the average number of antecedents of the three rules obtained using Continuous Re-RX was considerably lower than that obtained using C-MLP2LN and SSV.

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. We will attempt to extract more reliable diagnostic rules for diagnosis of Thyroid dysfunction in the future using updated Thyroid datasets that include TRAb and other important clinical and laboratory findings.

References

[1] Li LN, Ouyang JH, Chen HL, Liu DY. A computer aided diagnosis system for thyroid disease using extreme learning machine. J Med Syst 2012;36:3327–37.
[2] Chen HL, Yang B, Wang G, Liu J, Chen YD, Liu DY. A three-stage expert system based on support vector machines for thyroid disease diagnosis. J Med Syst 2012;36:1953–63.
[3] Dogantekin E, Dogantekin A, Avci D. An expert system based on Generalized Discriminant Analysis and Wavelet Support Vector Machine for diagnosis of thyroid diseases. Expert Syst Appl 2011;38:146–50.
[4] Dogantekin E, Dogantekin A, Avci D. An automatic diagnosis system based on thyroid gland: ADSTG. Expert Syst Appl 2010;37:6368–72.
[5] Pasi L. Similarity classifier applied to medical data sets, 2004, 10 sivua, Fuzziness in Finland '04. International conference on soft computing, Helsinki. Estonia: Finland & Gulf of Finland & Tallinn; 2004.
[6] Temurtas F. A comparative study on thyroid disease diagnosis using neural networks. Expert Syst Appl 2009;36:944–9.
[7] Serpen G, Jiang H, Allred L. Performance analysis of probabilistic potential function neural network classifier. In: Proceedings of artificial neural networks in engineering conference, viool. 7; 1997. p. 471–6.
[8] Ozyilmaz L, Yildirim T. Diagnosis of thyroid disease using artificial neural network methods, ICONIP '02. In: Proceedings of the 9th International Conference on Neural Information Processing, vol. 4; 2002. p. 2033–6.
[9] Liu DY, Chen HL, Yang B, XE L, Li LN, Liu J. Design of an enhanced fuzzy k-nearest neighbor classifier based computer aided diagnostic system for thyroid disease. J Med Syst 2012;36:3243–4354.
[10] Ng SK, McLachlan GJ. Extension of mixture-of-experts networks for binary classification of hierarchical data. Artif Intell Med 2007;41:57–67.
[11] Chang WW, Yeh WC, Huang PC. A hybrid immune-estimation distribution of algorithm for mining thyroid grand data. Expert Syst Appl 2010;37:2066–71.
[12] Kodaz H, Özşen S, Arslan A, Güneş S. Medical application of information gain based artificial immune recognition system (AIRS): Diagnosis of thyroid disease. Expert Syst Appl 2009;36:3086–92.
[13] Keleş A, Keleş A. ESTDD: expert system for thyroid diseases diagnosis. Expert Syst Appl 2008;34:242–6.
[14] Duch W, Adamczak R, Grąbczewski K. A new methodology of extraction, optimization and application of crisp and fuzzy logical rules. IEEE Trans Neural Netw 2001;12:277–306.
[15] Bologna G. A model for single and multiple knowledge based networks. Artif Intell Med 2003;28:141–63.
[16] Abe S, Thawonmas R. A fuzzy classifier with ellipsoidal regions. IEEE Trans Fuzzy Syst 1997;5:358–68.
[17] Abe S, Thawonmas R, Kayama M. A fuzzy classifier with ellipsoidal regions for diagnosis problems. IEEE Trans Syst Man Cybern C Appl Rev 1999;29:140–9.
[18] Sharaf-El-Deen DA, Moawad IF, Khalifa ME. A new hybrid case-based reasoning approach for medical diagnosis systems. J Med Syst 2014;38:9–19.
[19] Falco ID. Differential evolution for automatic rule extraction from medical databases. Appl Soft Comput 2013;13:1265–83.
[20] Zhu P, Hu Q. Rule extraction from support vector machines based on consistent region covering reduction. Knowl Based Syst 2013;42:1–8.
[21] Napierala K, Stefanowski J. BRACID: a comprehensive approach to learning rules from imbalanced data. J Intell Inf Syst 2012;39:335–73.
[22] Setiono R, Baesens B, Mues C. Recursive neural network rule extraction for data with mixed attributes. IEEE Trans Neural Netw 2008;19:299–307.
[23] University of California, Irvine Learning Repository. ⟨http://archive/ics.uci/edu/m/⟩ [accessed 10.10.15].
[24] Setiono R. Generating concise and accurate classification rules for breast cancer diagnosis. Artif Intell Med 2000;18:205–19.
[25] Setiono R. Extracting rules from pruned neural networks for breast cancer diagnosis. Artif Intell Med 1996;8:37–51.
[26] Setiono R, Liu H. Symbolic representation of neural networks. Computer 1996;29:71–7.
[27] Quinlan JR. C4.5: programs for machine learning. Morgan Kaufmann Series in Machine Learning. San Mateo, CA: Morgan Kaufman, Inc.; 1993.
[28] Setiono R. A penalty-function approach for pruning feedforward neural networks. Neural Comput 1997;9:185–204.
[29] Duch W, Adamczak R, Grąbczewski K. Extraction of logical rules from backpropagation networks. Neural Process Lett 1998;7:1–9.
[30] Hayashi Y, Fujisawa S. Strategic approach for the multiple-MLP ensemble Re-RX Algorithm. In: Proceedings of the 2015 International Joint Conference on Neural Networks (IJCNN), IEEE; 2015. p. 669–76.
[31] Nolan JP, Tarsa NJ, Dibenedetto G. Case-finding for unsuspected thyroid disease: costs and health benefits. Am J Clin Pathol 1985;83:346–55.
[32] Lan J, Hu MY, Patuwo E, Zhang GP. An investigation of neural network classifiers with unequal misclassification costs and group sizes. Decis Support Syst 2010;48:582–91.
[33] He H, Garcia E. Learning from imbalanced data. IEEE Trans. Knowl. Data Eng 2009;21:1263–84.
[34] Weiss G. Mining with rarity: a unifying framework. SIGKDD Explor 2004;6:7–19.
[35] Marqués AI, García V, Sánchez JS. On the suitability of resampling techniques for the class imbalance problem in credit scoring. J Oper Res Soc 2013;64:1060–70.
[36] Salzberg SL. On comparing classifiers: pitfalls to avoid and a recommended approach. Data Min Knowl Discov 1997;1:317–28.

[37] Mette AN, Torquil W, Bryan C, Laszlo H, Steen JB, Ase KR, Ulla FR, Jakob BB. Exploring the experiences of people with hypo- and hyperthyroidism. Qual Health Res 2014;25:945–53.
[38] Bahn RS, Burch HB, Cooper DS, Garber JR, Greenlee MC, Klein I, Laurberg P, McDougall IR, Montori VM, Rivkees SA, Ross DS, Sosa JA, Stan MN. Hyperthyroidism and other causes of thyrotoxicosis: management guidelines of the American Thyroid Association and American Association of Clinical Endocrinologists. Thyroid 2011;21:593–643.
[39] Jonklaas J, Bianco AC, Bauer AJ, Burman KD, Cappola AR, Celi FS, Cooper DS, Kim BW, Peeters RP, Rosenthal MS, Sawka AM. Guidelines for the treatment of hypothyroidism. Thyroid 2014;24:1670–751.
[40] ⟨http://www.japanthyroid.jp/en/guidelines.html⟩ [accessed 11.11.2015].