Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

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

Use of a Recursive-Rule eXtraction algorithm with J48graft to achieve highly accurate and concise rule extraction from a large breast cancer dataset

Yoichi Hayashi*, Satoshi Nakano

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

ARTICLE INFO

Article history:
Received 13 October 2015
Received in revised form
23 November 2015
Accepted 21 December 2015
Available online 15 February 2016

Keywords:
Breast cancer diagnosis
Rule extraction
Re-RX algorithm
J48graft
C4.5

ABSTRACT

To assist physicians in the diagnosis of breast cancer and thereby improve survival, a highly accurate computer-aided diagnostic system is necessary. Although various machine learning and data mining approaches have been devised to increase diagnostic accuracy, most current methods are inadequate. The recently developed Recursive-Rule eXtraction (Re-RX) algorithm provides a hierarchical, recursive consideration of discrete variables prior to analysis of continuous data, and can generate classification rules that have been trained on the basis of both discrete and continuous attributes. The objective of this study was to extract highly accurate, concise, and interpretable classification rules for diagnosis using the Re-RX algorithm with J48graft, a class for generating a grafted C4.5 decision tree. We used the Wisconsin Breast Cancer Dataset (WBCD). Nine research groups provided 10 kinds of highly accurate concrete classification rules for the WBCD. We compared the accuracy and characteristics of the rule set for the WBCD generated using the Re-RX algorithm with J48graft with five rule sets obtained using 10-fold cross validation (CV). We trained the WBCD using the Re-RX algorithm with J48graft and the average classification accuracies of 10 runs of 10-fold CV for the training and test datasets, the number of extracted rules, and the average number of antecedents for the WBCD. Compared with other rule extraction algorithms, the Re-RX algorithm with J48graft resulted in a lower average number of rules for diagnosing breast cancer, which is a substantial advantage. It also provided the lowest average number of antecedents per rule. These features are expected to greatly aid physicians in making accurate and concise diagnoses for patients with breast cancer.

1. Introduction

Cancer remains a devastating health problem in the United States, with nearly 1.7 million new cases and 600,000 estimated in 2015. An estimated 28.6% (810,170) of new cancer cases among females involve breast cancer, making it the most frequently diagnosed type of new cancer among women [1]. Therefore, breast cancer diagnosis has become an increasingly important issue in the medical field.

The American Cancer Society estimated that more than 230,000 cases of invasive and nearly 65,000 cases of noninvasive breast cancer were diagnosed in the United States in 2013 [2], and that nearly 40,000 of these cases were fatal. However, recent improvements in breast cancer survival have been evident; this improvement likely involves a variety of factors, including a higher rate of screening mammography, which allows the diagnosis and treatment of breast cancer at earlier, more treatable stages, and new classes of chemotherapeutic agents. However, despite these improvements, a number of factors continue to be associated with poorer survival in all stages of breast cancer [3].

Breast cancer, which globally is the second most common type of cancer and the fifth most common cause of cancer death, is the most common type of cancer among females, with an incidence more than twice those of colorectal and cervical cancers, and a 25% higher mortality rate than that of lung cancer.

However, great progress in detecting breast cancer at an earlier stage is being made. Early diagnosis of breast cancer requires an accurate and reliable procedure that allows physicians to distinguish between benign and malignant tumors [4]; therefore, expert systems and artificial intelligence techniques are increasingly being developed to improve diagnostic capabilities. These automatic diagnostic systems can help avoid human errors made in the course of diagnosis, and allow the data to be examined in less time and greater detail.

During breast cancer diagnosis, physicians form an opinion about the condition of a tumor and decide whether it is benign or malignant

* 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).

http://dx.doi.org/10.1016/j.imu.2015.12.002

based on an examination of the patients' symptoms. Currently, physicians (breast surgeons) carefully follow American Cancer Society guidelines [2] or other national standards for the early detection of the breast cancer. Breast cancer diagnosis varies depending on the age of the patient, and typical methods include mammography and clinical breast examination (CBE), fine needle aspiration (FNA) cytology, ultrasonography-guided vacuum-assisted core needle biopsy (CNB), and, for patients at high risk, magnetic resonance imaging.

However, even experienced physicians can sometimes delay making a definitive diagnosis. Therefore, to assist physicians in the diagnosis of breast cancer, a highly accurate computer-aided diagnostic system is necessary.

In effort to increase the diagnostic accuracy and processing of increasingly large amounts of tumor data and information, a number of researchers have turned to machine learning approaches and data mining, a tool that allows the discovery of knowledge behind large scale data that has been shown to be highly applicable in real world settings. Data mining and machine learning have been incorporated into a computer-aided diagnostic system for breast cancer since 1995 [5].

In 1996, Setiono proposed a method based on a neural network (NN) pruning technique to extract concrete classification rules for the Wisconsin Breast Cancer Dataset (WBCD) [6,7]. The idea underlying the approach was to take advantage of the expressive power provided by sets of IF–THEN rules; this is an extremely effective diagnostic technique in the medical domain.

The WBCD is the result of efforts made at the University of Wisconsin Hospital to accurately diagnose breast masses based solely on an FNA test in 1992. This technology is still used today and known as FNA cytology. FNA cytology has been used extensively over the years in the diagnosis of breast lesions.

Diagnostic accuracy can be achieved through a multidisciplinary consultation, combining FNA cytology results with CBE and imaging modalities such as mammography and ultrasonography (triple assessment). The diagnostic value of FNA cytology improves with the immediate on-site evaluation of specimens. Immediate cytologic diagnosis in real time is cost-effective and allows patients with benign diseases to be given immediate reassurance; it also allows the quick planning of management for patients with malignant or suspicious lesions [8].

In 1999, a neuro-fuzzy approach for breast cancer diagnosis was proposed by Nauck and Kruse [9]. Although their approach was based on fuzzy clustering rather than rule extraction, their research was the first to provide concise fuzzy rules and obtain results using 10-fold cross validation (CV) [10]. Therefore, in Section 4, we compare the results from the present study with those of Nauck and Kruse [9] and investigate the performance of their extracted rules.

Also in 1999, Peňa-Reys and Sipper proposed a fuzzy-genetic diagnostic approach [11] for the WBCD. Their approach exhibited two promising characteristics: first, it attained high classification performance; second, the resulting systems involved only a few simple rules, and was therefore human-interpretable.

As a result, their approach confirmed that data mining technologies could be successfully implemented in cancer prediction, allowing traditional breast cancer diagnosis to be transformed into a classification problem in the data mining domain. A classifier was then devised to categorize tumors in existing datasets as benign or malignant. Then, based on an evaluation of the classifier and the historical tumor data, new tumors could be predicted [12].

Breast cancer diagnosis can be formulated as a two-class classification problem. Classification is one of the most frequently faced tasks in many different fields, and is of paramount importance among physicians in decision making regarding diagnosis [13].

For the diagnosis of breast cancer with high classification accuracy, numerous types of artificial intelligence, computational intelligence, and other techniques have been investigated, including neuro-fuzzy systems [9,14], NNs [4,7,15–22], sequential covering algorithm [23], support vector machines (SVMs) [4,24–31], linear discriminant analysis (LDA) [32], fuzzy clustering [33], artificial immune systems [34–36], case-based reasoning [37], mixture of experts [38], differential evolution [13], artificial metaplasticity algorithm [39], fuzzy-rough nearest neighbor classifier [40], HMM-fuzzy approach [41], and fuzzy entropy-based feature selection [42].

However, most of the current diagnostic methods for breast cancer are black-box models that are unable to satisfactorily reveal hidden information in the data that typically plays a key role in providing a quality medical diagnosis.

For example, even though a method may correctly assign an instance to a group, it still does not provide users with information regarding the reasons why the item was classified in a specific way. Therefore, algorithms that provide insight into the rationale behind their behavior are highly sought, and an increasing amount of research is being devoted to the user-friendliness of systems and the self-explicability of their behavior [13].

Rule extraction is a powerful method of data mining that provides explanation capabilities, knowledge discovery, and knowledge acquisition; therefore, it is becoming increasingly popular. However, algorithms for rule extraction should meet several crucial requirements for practical use. Extracted rules need to be simple and human-interpretable, and must be able to discover highly accurate knowledge in the medical domain.

In previous studies, some researchers have extracted Boolean rules from NNs in an attempt to gain increased interpretability [7,15,43]. The results of these studies were encouraging, as the use of Boolean rules led to good performance, a reduced number of rules, and relevant input variables. However, because these systems use Boolean rules, they are not capable of continuous rules.

The Recursive-Rule eXtraction (Re-RX) algorithm, originally intended to be a rule extraction tool, was recently developed by Setiono et al. [44]. Re-RX provides a hierarchical, recursive consideration of discrete variables prior to analysis of continuous data and can generate classification rules from NNs that have been trained on the basis of both discrete and continuous attributes.

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.

To achieve both conciseness and high accuracy of extracted rules while simultaneously maintaining the good framework of the Re-RX algorithm, we recently proposed supplementing the Re-RX algorithm with J48graft, a class for generating a grafted C4.5 decision tree (hereafter Re-RX with J48graft) [45].

The J48graft [46] is the result of the C4.5A [47] algorithm being implemented in open source data mining software, which was introduced by Webb and referred to as the "all-tests-but-one partition (ATBOP)" [47].

In Re-RX with J48graft, J48graft [46] is employed to form decision trees in a recursive manner, while multi-layer perceptrons (MLPs) are trained using backpropagation (BP), which allows pruning [6], thereby generating more efficient MLPs for highly accurate rule extraction.

In contrast to these black-box models, Re-RX with J48graft not only provides extremely high classification, but also can be 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 thus often preferred by physicians and clinicians.

Typically, the accuracy of extracted classification rules is judged by the number of test or training samples that have been correctly classified, while interpretability is judged by the complexity of the model; more specifically, the number of extracted rules and the average number of antecedents they contain.

Current data mining research, especially high performance classifier research, seems focused only on predictive accuracy. Rule extraction is a technique that attempts to find compromise between both requirements by building a simple rule set that mimics how the well-performing complex model (black-box) makes it decisions for physicians, breast surgeons, and pathologists. Furthermore, more attention needs to be paid to the relationship between the expressive power and the quality of the extracted rules.

Many types of rules have been suggested in the literature. Propositional rules take the form of IF–THEN expressions, where clauses are defined in propositional or fuzzy logic. The trade-off between the number of rules and the average number of antecedents also needs to be balanced. Removal of redundant and irrelevant antecedents enhances the expressive power and the quality of the extracted rules, which are more concise and suitable for medical decision making. Ideally, both high accuracy and interpretability should be achieved simultaneously [48].

Previous studies [7,9,13,15–18,22,24,30,36] have reported results and provided the number of extracted classification rules in the diagnostic approach for the WBCD, and concrete rule sets have also been reported [7,9,11,13,15–17,24,30,36].

The objective of the present study was to achieve highly accurate, concise, and interpretable classification rules for breast cancer diagnosis. However, in this paper, the target dataset for rule extraction was a medical dataset, i.e. the WBCD; therefore, the focus was on decreasing the number of extracted rules and the average number of antecedents. To extract concise rules, Re-RX with J48graft [45], which is better suited for achieving concise as opposed to accurate medical rules, was employed. The WBCD was obtained from the University of California Irvine Machine Learning Repository [49].

The most important aim of this study was to improve the interpretability of extracted rules for physicians because the competition for achieving better accuracy for academic medical datasets has appeared to plateau, and unless diagnostic accuracy can be substantially improved more than just a few percentage points, no significant contributions will be made to medical informatics.

In Sections 4.2.1–4.2.6, six kinds of rule sets extracted from WBCD are explained. The interpretability and conciseness of these extracted rules vary widely.

Another aim of this study was to validate the generalization capability of Re-RX with J48graft using 10 runs of 10-fold CV. If generalization capability is sufficiently high, Re-RX with J48graft can be used to extract diagnostic rules for different kinds of breast cancer datasets with different cytopathology characteristics.

Nine research groups [7,9,11,13,15–17,24,30,36] provided 10 kinds of highly accurate concrete classification rule sets for the WBCD. We compared the accuracy and characteristics of the rule set for the WBCD obtained by Re-RX with J48graft with four rule sets obtained using 10 runs of 10-fold CV [10]; these five rule sets obtained by k-fold CV have been previously reported [9,17,24,30].

2. Theory

2.1. Re-RX algorithm

The Re-RX algorithm generates classification rules from both continuous and discrete datasets. It produces hierarchical rules, applying different rule conditions for discrete and continuous attributes, such that only the rules lowest in the hierarchy contain continuous attributes. Here, although the proposed algorithm can readily handle multiple groups, two-group classification problems are considered exclusively. The algorithm structure and functioning are described as follows.

```
Algorithm Re-RX (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 [6] an NN using dataset S, including 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 $D'=\phi$ (empty), generate an axis hyperplane to split the
  samples in S' according to the values of the continuous
  attributes, $C'$, then stop.
Otherwise, use only the discrete attributes, $D'$, to generate the
  set of classification rules, R, for dataset S'.
4. For each rule, $R_i$, that 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 satisfy the condition of rule
  $R_i$, and let $D_i$ be the set of discrete attributes that do not appear
  in rule condition $R_i$.
• If $D_i=\phi$, then generate hyperplane to split the samples in $S_i$
  according to the values of their continuous attributes, $C_i$, then
  stop.
• Otherwise, call Re-RX ($S_i$, $D_i$, $C_i$).
```

Assuming a suitable pruning rate, Step 1 can employ a variety of NN training and pruning methods. Although the Re-RX algorithm makes no assumptions regarding the NN architecture, we have focused on BPNNs with a single hidden layer, allowing universal approximation.

The percentage of samples covered by a rule defines its support, and Step 4 assesses both the rule support and the corresponding error rate. The rule subspace is further partitioned if the error rate is above a threshold value, δ_2, and the support equals the approximate maximum threshold value, δ_1. If discrete attributes are absent from the rule conditions, subdivision is achieved by recursively calling Re-RX or by producing a separate axis hyperplane incorporating only the continuous data attributes.

2.2. J4.8

J4.8 [50] is a Java-implemented version of C4.5 [51], an advanced version of the ID3 algorithm developed by Quinlan [52]. The decision trees generated by C4.5 are used for classification; therefore, this algorithm is typically described as a statistical classifier. C4.5 performs very similarly to ID3, except that it determines the best target attribute using the gain ratio. Also, in contrast to ID3, C4.5 has the improved ability to handle numerical attributes by creating a threshold, and then splitting the data into those whose attribute value is either greater, or less than or equal to, that threshold. This algorithm also has the ability to handle attributes with variable cost. Finally, C4.5 can prune the decision tree after its creation, which reduces its size and thereby saves both time and memory.

2.3. J48graft

The concept of tree grafting is based on the desire to discard the "simplest is best" method for selecting a good tree. In contrast, in tree grafting, the focus is on the fact that similar objects tend to have the highest probability of belonging to the same class. In other words, if the final result is a better classification model, the need to yield more complex trees is eliminated.

Grafting is a post-process that can be readily applied to decision trees. Its main objective is to reclassify regions of an instance space where no training data exists or where there is only misclassified data, and as a result, to decrease prediction error. Grafting identifies the best-suited cuts of existing leaf regions and then branches out to create new leaves with classifications that differ from the original. In this process, the tree becomes more complex naturally. However, only branching that does not introduce classification errors in data that has already been correctly classified is considered, ensuring that the new tree reduces errors.

Webb introduced the C4.5A algorithm referred to as the "all-tests-but-one partition (ATBOP)," which is a more efficient method for evaluating potentially supporting evidence [47]. The ATBOP region of a leaf is formed by removing all the enclosing decision surfaces. Using ATBOP allows a reduction in computational requirements because the only set of training data considered for each leaf is that from the ATBOP region. The J48graft is the result of the C4.5A algorithm being implemented in open source data mining software known as the Waikato Environment for Knowledge Analysis (Weka) [50].

Pruning is a process that can be thought of as the opposite of grafting because it aims to reduce rather than increase the complexity of a decision tree while retaining good predictive accuracy. Surprisingly, Webb [53] concluded that, either despite or possibly because of the fact they are opposites, pruning and grafting work well in parallel. Grafting takes instances outside the analyzed leaf (global information) into account, while pruning only looks at instances within the analyzed leaf (local information). In this way, they seem to complement each other. In most cases, using both grafting and pruning on a decision tree yields a lower prediction error that using them separately [53].

3. Materials and methods

3.1. Re-RX algorithm with J48graft

To enhance the accuracy and conciseness of classification rules, we propose replacing the conventional Re-RX algorithm, which uses C4.5 as a decision tree [52], with Re-RX with J48graft. Concepts in the conventional pruning used in J4.8 and that used in J48graft [46] both contrast and complement each other. We believe that the performance of the Re-RX algorithm [44] is greatly affected by the decision tree. In consideration of the grafting properties in J48graft, our idea is to use the grafting concepts in the Re-RX algorithm to enhance the accuracy and conciseness of the extracted rules. Therefore, we replace J4.8 with J48graft in the Re-RX algorithm. We also expect that Re-RX with J48graft will generate much more accurate and concise classification rules.

In summary, we frequently employ J48graft in Re-RX with J48graft [45] to form decision trees in a recursive manner, while we train MLPs using BP, which allows pruning [6] and therefore generates more efficient MLPs for rule extraction. The schematic overview of the Re-RX with J48graft is shown in Fig. 1.

3.2. Data description and experimental setup

In this study, we conducted rule extraction experiments for the WBCD [49], a dataset that comprises 699 instances taken from needle aspirates from the patient's breast. Among these cases, 458 were classified as benign class and the remaining 241 as malignant. A total of 16 instances had missing values, and for the purposes of this study, all missing values were replaced by the mean of the attributes. The remaining 683 (444 benign, 239 malignant) cases were randomly divided into a training set consisting of 222 benign and 119 malignant cases. The remaining cases represented the test set.

Although the presence of a breast mass is a cause for concern, it does not always indicate a malignant cancer. FNA of breast masses is a non-invasive, non-traumatic, and cost-effective diagnostic test that provides information necessary for evaluating malignancy. The WBCD is the result of efforts made at the University of Wisconsin Hospital for accurately diagnosing breast masses based solely on an FNA test from 1992. Nine visually-assessed characteristics of an FNA sample considered relevant for diagnosis were identified, with each characteristic assigned an integer value between 1 and 10 (with 1 being the closest to benign and 10 being the most anaplastic).Therefore, each record in the dataset has nine attributes (Table 1). For WBCD preprocessing, the value of each input attribute was normalized in a range between 0 and 1.0.

3.3. FNA and CNB

The WBCD was developed with the aim of enabling accurate diagnosis of breast masses based solely on an FNA test. Each sample consisted of nine attributes of FNA samples considered relevant for diagnosis. The role of FNA compared with CNB is as follows.

Although FNA is the established cell collection biopsy technique for breast masses, a shift toward the use of CNB with image guidance has been occurring due to concerns over low accuracy and high rates of inadequacy in FNA specimens.

Nagar et al. [54] concluded that FNA is a highly accurate biopsy technique in palpable breast lesions when performed correctly, and that a majority of patients receive definitive therapy on the basis of their cytology. FNA typically has comparable predicative values with, as well as cumulative treatment costs lower than, CNB, and thus FNA is still considered to be the first-line pathologic investigation for palpable breast lesions.

In fact, no absolute false-positives or false-negatives have been reported in patients receiving either CNB or FNA, and no significant difference was evident in their predictive values. FNA without ultrasound is also substantially less costly than CNB with ultrasound.

The current era is marked by a constant striving to find ways to decrease the cost of health care delivery without compromising quality, and thus performing FNA for palpable breast lesions leads to great cost savings for society.

Therefore, we believe that results from FNA are an appropriate candidate input attribute for constructing diagnostic breast cancer datasets

4. Results

4.1. Experimental results

In order to guarantee the validity of the results, we used k-fold CV [10] to evaluate the classification rule accuracy of test datasets. The k-fold CV method is widely applied by researchers to minimize the bias associated with random sampling.

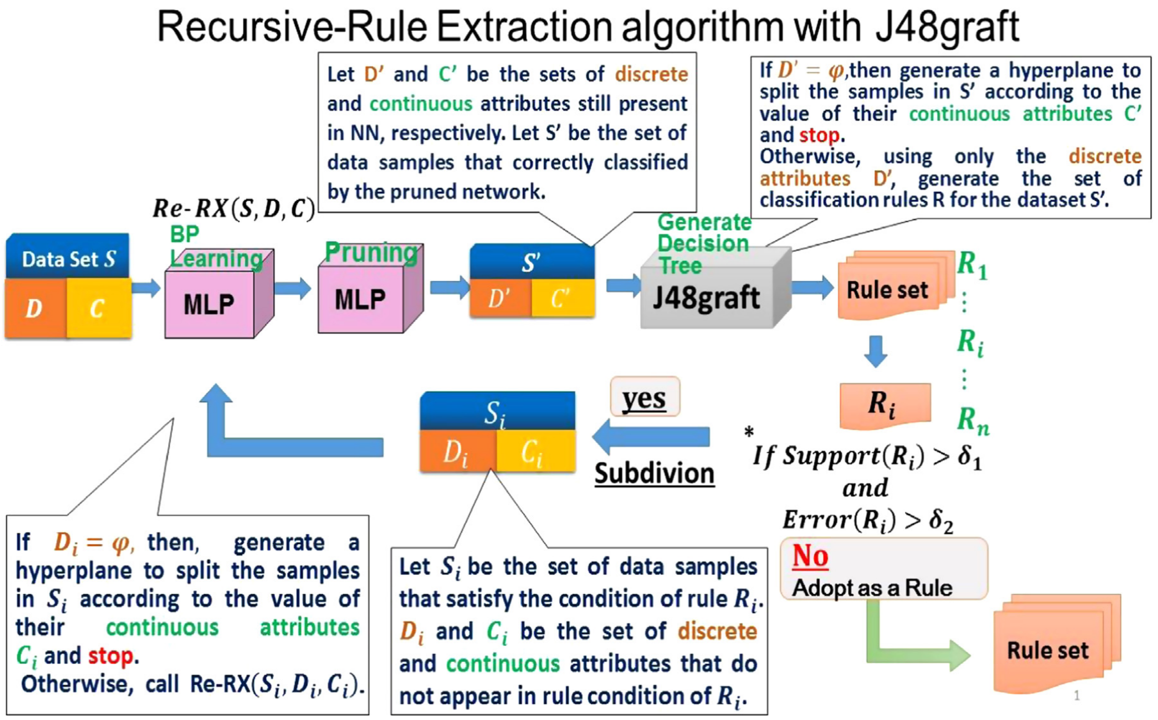
Recursive-Rule Extraction algorithm with J48graft
Let D′ and C′ be the sets of discrete and continuous attributes still present in NN, respectively. Let S′ be the set of data samples that correctly classified by the pruned network.
If D′ = φ, then generate a hyperplane to split the samples in S′ according to the value of their continuous attributes C′ and stop.
Otherwise, using only the discrete attributes D′, generate the set of classification rules R for the dataset S′.
Re-RX(S, D, C)
Data Set S
D
C
BP Learning
MLP
Pruning
MLP
S′
D′
C′
Generate Decision Tree
J48graft
Rule set
R1
Ri
Rn
Ri
yes
Si
Di
Ci
Subdivion
* If Support(Ri) > δ1 and Error(Ri) > δ2
No
Adopt as a Rule
Rule set
If Di = φ, then, generate a hyperplane to split the samples in Si according to the value of their continuous attributes Ci and stop.
Otherwise, call Re-RX(Si, Di, Ci).
Let Si be the set of data samples that satisfy the condition of rule Ri.
Di and Ci be the set of discrete and continuous attributes that do not appear in rule condition of Ri.
1

Fig. 1. Schematic overview of the Recursive-Rule eXtraction algorithm with J48graft.

Table 1
The nine attributes of breast cancer data from the Wisconsin Breast Cancer Dataset (WBCD).

Label	Attribute	Abbreviation	Domain
F_1	Clump thickness	CT	1-10
F_2	Uniformity of cell size	UCSI	1-10
F_3	Uniformity of cell shape	UCSH	1-10
F_4	Marginal adhesion	MA	1-10
F_5	Single epithelial cell size	SECS	1-10
F_6	Bare nuclei	BN	1-10
F_7	Bland chromatin	BC	1-10
F_8	Normal nucleoli	NN	1-10
F_9	Mitoses	M	1-10

WBCD: Wisconsin Breast Cancer Dataset.

Table 2
Performance of the Recursive-Rule eXtraction (Re-RX) algorithm with J48graft for the WBCD (10 runs of 10-fold cross validation).

WBCD	TR ACC (%)	TS ACC (%)	# Rules	Ave. # ante.	AUC	TR ACC (SD)	TS ACC (SD)
Re-RX with J48graft	96.30	95.80	**4.8**	**1.69**	0.959	1.80	1.65

Re-RX: Recursive-Rule eXtraction, WBCD: Wisconsin Breast Cancer Dataset, TR: training dataset, TS: testing dataset, ACC: accuracy, Ave. # ante.: average number of antecedents, AUC: area under the receiver operating characteristic curve, SD: standard deviation.

Table 3
Performance of the Re-RX algorithm for the WBCD (10 runs of 10-fold cross validation).

WBCD	TR ACC (%)	TS ACC (%)	# Rules	Ave. # ante.	AUC	TR ACC (SD)	TS ACC (SD)
Re-RX with J48graft	95.60	95.09	6.6	2.91	0.947	1.80	1.45

Re-RX: Recursive-Rule eXtraction, WBCD: Wisconsin Breast Cancer Dataset, TR: training dataset, TS: testing dataset, ACC: accuracy, Ave. # ante.: average number of antecedents, AUC: area under the receiver operating characteristic curve, SD: standard deviation.

We trained the WBCD using Re-RX with J48graft and the average classification accuracies of 10 runs of 10-fold CV for the training dataset (TR ACC), the average classification accuracies of 10 runs of 10-fold CV for the test dataset (TS ACC), the number of extracted rules (# rules), the average number of antecedents (Ave. # ante.), and the area under the receiver operating characteristic (ROC) curve (AUC) [54] for the test dataset (Table 2). In this paper, the AUC was used as an appropriate performance evaluator because it does not include class distribution or misclassification costs [55]. In addition, “ELSE” (the so-called “default rule”) appears in the extracted rule set as one rule.

Numerous types of rules have been suggested in the literature from the perspective of the expressive power of extracted rules, including propositional rules, which take the form of IF–THEN expressions and clauses are defined in propositional logic, and M-of-N rules. Breaking from traditional logic, fuzzy rules allow partial truths instead of Boolean true/false outcomes.

Even if all types of rules are considered, the consensus is that no matter how they are defined, an ideal measure has yet to be developed; therefore, “what is a concise and/or interpretable rule?” remains a difficult question to answer.

In order to more precisely answer this question, we attempted to develop a “rough indicator” of conciseness by comparing the average number of antecedents from extracted rules generated using a variety of techniques.

Regarding the complexity of Re-RX with J48graft, it took about 5 seconds to train the WBCD using a standard workstation computer (3.1 GHz Intel Xeon E5-2687W, 3.5 GHz Turbo, 25 MB Cache; 64 GB RAM; 512 GB DDR3 System memory) and about 45 seconds for 10-fold CV. The testing time was negligible.

For reference, the performance by the original Re-RX algorithm [44], i.e. Re-RX with C4.5, is shown in Table 3.

Table 4
Performance of previous rule extraction algorithms for the Wisconsin Breast Cancer Dataset (WBCD).

Rule extraction method [validation method]	TR ACC (%)	TS ACC (%)	# Rules	Rule set	Ave. # ante	Year [Ref.]
Neural network pruning [train: 50%-test: 50%]	96.29–97.71	93.12–96.56	6	Yes	2.83	1996 [7]
NeuroLinear [train: 50%-test: 50%]	97.14	94.27	2	Yes	3.0	1997 [15]
NEFCLASS [10CV]	–	95.06	2 (Fuzzy rule)	Yes	5.5	1999 [9]
Fuzzy-genetic approach [train: 50%-test: 50%]	–	97.8	6 (Fuzzy rule)	Yes	2.33	1999 [11]
NeuroRule [train: 50%-test: 50%]	–	97.95	6	Yes	3.5	2000 [16]
C-MLP2LN [10CV]	–	99.0	6	Yes	3.67	2001 [17]
SSV [10CV]	–	96.3 ± 0.2	4	Yes	2.0	2001 [17]
MINERVA [10CV]	–	94.52 ± 1.51	4.20	–	3.33	2008 [23]
NeuroLinear+GRG [10CV]	–	95.96	2	–	–	2008 [18]
CLONALG [Max. ACC]	–		4	Yes	3.5	2012 [36]
GASVM [4CV]	–	94.6	4	Yes	2.5	2013 [30]
DEREx [Max. ACC]	98.05	100	3	Yes	5.0	2013 [13]
DIMLP-B [10 × 10CV]	98.0 ± 0.1	97.4 ± 0.2	12.5	–	2.7	2015 [24]
DIMLP-B [10 × 10CV]	100.0 ± 0.0	96.9 ± 0.3	25.2	–	3.6	2015 [24]
QSVM-L [10 × 10CV]	95.2 ± 0.0	95.6 ± 0.3	13.2	–	3.0	2015 [24]
QSVM-P3 [10 × 10CV]	92.3 ± 0.0	92.9 ± 0.3	22.7	–	3.5	2015 [24]
QSVM-G [10 × 10CV]	97.20 ± 0.0	97.5 ± 0.2	12	Yes	2.67	2015 [24]
Re-RX algorithm with J48graft [10 × 10CV]	96.30 ± 1.80	95.80 ± 1.65	4	Yes	1.75	Present paper

Data are expressed as mean ± standard deviation.
Re-RX: Recursive-Rule eXtraction, WBCD: Wisconsin Breast Cancer Dataset, TR: training dataset, TS: testing dataset, ACC: accuracy, Ave. # ante.: average number of antecedents, 10CV: 10-fold cross validation, 4CV: 4-fold cross validation, 10 × 10CV: 10 runs of 10-fold cross validation.

Comparing the number of rules and the average number of antecedents, Re-RX with J48graft provided markedly fewer rules and average number of antecedents.

The accuracy for the training and test datasets, the number of extracted rules, and the average number of antecedents obtained by classification rule algorithms proposed in previous papers are shown in Table 4.

Table 4 shows rule extraction algorithms proposed for the WBCD since 1996. We conducted a comparison in relation to the performance of rule extraction algorithms for the WBCD based on unified statistical validation using five rule sets obtained by k-fold CV [9,17,24,30]. This is discussed in further detail in Sections 4.2 and 4.3.

4.2. Extracted rule sets obtained by previous and the present algorithms

In this section, we demonstrate the six kinds of rule sets extracted from the WBCD by NEFCLASS, C-MLP2LN, SSV, GASVM, QSVM-G, and Re-RX with J48graft. All attributes of the WBCD appearing in the extracted rules are as abbreviated as follows: CT: Clump thickness; UCSI: Uniformity of cell size; UCSH: Uniformity of cell shape; MA: Marginal adhesion; SECS: Single epithelial cell size; BN: Bare nuclei; BC: Bland chromatin; NN: Normal nucleoli; and M: Mitoses.

4.2.1. Extracted rule set for the WBCD by NEFCLASS [9]

R_1: IF CT is *large* AND UCSH is *large* AND MA is *large* AND BN is *large* AND BC is *large* AND NN is *large* THEN Malignant.
R_2: IF UCSH is *small* AND MA is *small* AND BN is *small* AND BC is *small* AND NN is *small* THEN Benign.

4.2.2. Extracted rule set for the WBCD by C-MLP2LN [17]

R_1: IF CT < 6 AND UCSH < 3 AND BC < 8 THEN Malignant
R_2: IF CT < 9 AND MA < 4 AND BN < 2 AND BC < 5 THEN Malignant
R_3: IF CT < 10 AND UCSH < 4 AND MA < 4 AND BN < 3 THEN Malignant
R_4: IF CT < 7 AND UCSH < 9 AND MA < 3 AND BN ∈ [4,9] AND BC < 4
THEN Malignant
R_5: IF CT ∈ [3,4] AND UCSH < 9 AND MA < 10 AND BN < 6 AND BC < 8
THEN Malignant
ELSE Benign

4.2.3. Extracted rule set for the WBCD by SSV [17]

R_1: IF MA > 2.5 AND BC > 2.5 THEN Malignant
R_2: IF MA > 2.5 AND BN > 3.5 AND BC THEN Malignant
R_3: IF UCSI > 5.5 AND MA < 2.5 AND BC > 1.6 THEN Malignant
ELSE Benign

4.2.4. Extracted rule set for the WBCD by GASVM [30]

R_1: IF CT < 7.085 AND UCSH < 7.908 AND SECS < 9.76
AND BC < 6.064
THEN Benign
ELSE: Malignant
R_2: IF UCSH < 7.70 AND BN < 9.41 AND BC < 6.12 AND M < 7.43
THEN Malignant
ELSE: Benign

4.2.5. Extracted rule set for the WBCD by QSVM-G [24]

R_1: (CT < 9.99462) AND (UCSH < 3.98945) AND (BN < 6.92955) AND
(BC < 6.95644) AND (NN < 9.94272) THEN Benign
R_2: (CT < 6.99821) AND (UCSI < 5.96902) AND (SECS < 4.96611) AND (BN < 4.93843) AND (NN < 9.94272) THEN Benign
R_3: (UCSH > 2.97063) AND (BN > 4.93843) THEN Malignant
R_4: (CT > 4.96292) AND (UCSI > 3.9883) THEN Malignant
R_5: (UCSI > 4.97866) THEN Malignant
R_6: (CT > 2.98416) AND (BN > 6.92955) THEN Malignant
R_7: (CT > 5.98057) AND (UCSI > 2.99794) AND
(UCSH > 3.98945) THEN Malignant
R_8: (UCSH > 2.97063) AND (SECS > 4.96611) THEN Malignant
R_9: (NN > 8.964) THEN Malignant
R_{10}: (UCSI < 2.99794) AND (UCSH > 3.98945) AND
(SECS < 4.96611) THEN Benign
R_{11}: (CT < 2.98416) AND (UCSH < 4.94833) AND
(BN > 9.9531) THEN Benign

R_{12}: $(CT > 9.99462)$ AND $(UCSI < 2.99794)$ AND $(BN < 7.96198)$ Benign

4.2.6. Extracted rule set for the WBCD by Re-RX with J48graft [45]

R_1: IF $BN = 1$ THEN Benign
R_2: IF $CT \leq 4$ AND $1 < BN \leq 6$ THEN Benign
R_3: IF $CT \leq 4$ AND $BN > 6$ THEN Malignant
R_4: IF $CT > 4$ AND $BN > 1$ THEN Malignant

4.3. Comparisons

The experiment carried out on the WBCD resulted in the extraction by 10-fold CV of four rules for diagnosis with high accuracy (95.80%). These four rules were not only more accurate, but also more comprehensible in terms of the average number of antecedents per rule (1.75), which was considerably smaller than the algorithms described in Sections 4.2.1–4.2.5.

Two rules obtained by NEFCLASS consisted of 11 linguistic variables that demonstrated high expressive power and accuracy (95.6%) for the test dataset. However, the average number of antecedents was dependent on the definition of membership functions for the linguistic variables in the two rules. The average number of antecedents was largest (5.5) among the five algorithms discussed in Section 4.

Six rules obtained by C-MLP2LN achieved the highest accuracy (99.0%), but the average number of antecedents (3.67) was much larger than that (1.75) obtained in the present study.

Four rules obtained by SSV provided 96.3 ± 0.3% accuracy and a concise average number of 2.0 antecedents, which are approximately the same as the 95.80 ± 1.65% accuracy, but slightly larger than the average of 1.75 antecedents obtained in the present study.

Four rules obtained by GASVM provided 94.6% accuracy and an average of 2.5 antecedents, which are approximately the same as the 95.80 ± 1.65% accuracy, but slightly larger than the average of 1.75 antecedents obtained in the present study.

A total of 11 rules obtained by QSVM-G achieved high accuracy (97.5 ± 0.2), but there were such a large number of extracted rules that they were difficult to interpret. The average number of antecedents was 2.67.

Four rules obtained by Re-RX with J48graft achieved 95.80% accuracy, which is approximately the same as that obtained by SSV. The number of extracted rules was the same as that of the SSV, while the average number of antecedents was 1.75, which was slightly smaller than that of SSV (2.0), and less than half that of C-MLP2LN (3.67).

5. Discussion

5.1. Cytopathology interpretation of extracted rules by Re-RX with J48graft

In FNA cytology, the presence of BN represents the benignity of the cell [56]. It has also been reported that BN within a breast aspirate is generally indicative of a benign lesion [57]. CT is described as the number of layers of the smear sample, and is categorized as monolayered, monolayered and folding, or multilayered [56]. Mitoses is the process of nuclear division in cells that produces daughter cells which are genetically identical to each other and to the parent cell. Malignant cells tend to have higher mitotic activities compared with normal and benign cells. The absence of mitoses has also been reported as a cytologic finding in benign granular cell tumors [58].

Therefore, we believe that BN and CT, as well as mitoses, are important for diagnosing breast cancer and should be included as attributes in extracted rules. If a relatively large number of mitoses are present, mitoses should be included as an attribute in the extracted rules for a malignant diagnoses. However, out of all 683 WBCD samples, the number of mitoses in domain 1 is 563. This means that few mitoses samples are categorized to the high domain, and therefore has an extremely biased distribution. Consequently, mitoses is not included as an attribute of benign or malignant rules for the WBCD.

Although it is difficult, we believe that concise diagnostic rules for various kinds of breast cancer can be extracted from the WBCD to assist physicians.

5.2. Significance of highly accurate rule extraction by Re-RX with J48graft for the WBCD

Rule extraction algorithms proposed in previous papers often require the discretization of continuous attributes in the datasets. However, this discretization often causes non-negligible amounts of information to be lost in order to achieve highly accurate rule extraction. To resolve this problem, Setiono et al. [43] proposed the Re-RX algorithm for extracting rules for mixed datasets that consist of discrete—binary, categorical, and nominal—and continuous attributes.

Re-RX with J48graft is a concise rule extractor which is able to handle not only mixed datasets, but also discrete or continuous attribute datasets separately. The WBCD belongs to the latter case. Medical datasets often belong to the mixed dataset with multiple classes.

The clear advantage of Re-RX with J48graft over other algorithms is the fact that it provides physicians with explicit knowledge extracted from the WBCD in the form of IF–THEN rules. In fact, it can express rules to perform diagnosis much more clearly than its competitors, such as support vector machines and differential evolution.

Furthermore, Re-RX with J48graft can also perform highly accurate and concise rule extraction, since the achieved rules contain some of the WBCD attributes only; this can be seen as extremely supportive in regards to achieving accurate diagnoses. These rules are therefore proposed for use by physicians so that Re-RX with J48graft may aid the diagnosis of breast cancer and explanations regarding the reasons why a patient is believed to suffer or not to suffer from a given pathology. In this way, we aim to help physicians by providing them with more useful and accurate information.

As previously mentioned, highly accurate screening can potentially be used to diagnose breast cancer in massive numbers of patients. However, the diagnostic accuracy for the WBCD in the present study showed that the loss of vast amounts of misdiagnosed cases could be prevented by increasing the performance of rule extraction for the WBCD, even to a small extent.

6. Conclusions

Rule extraction is important for the diagnosis of breast cancer. An attractive feature of rule extraction is that it provides physicians with highly accurate, concise, and interpretable rules extracted from the WBCD. Of course, the knowledge acquired should never substitute that of the expert, but rather should be seen as a way to support appropriate decision making.

In this study, we employed Re-RX with J48graft for rule extraction from the WBCD. Re-RX with J48graft provides highly accurate, concise, and interpretable rules from the WBCD in the form of IF–THEN rules.

Compared with other rule extraction algorithms, Re-RX with J48graft results in a lower average number of rules for diagnosing breast cancer, which is a substantial advantage. Moreover, it also provides the lowest average number of antecedents per rule. These features are expected to provide information that is extremely useful to physicians. Of course, the opinions of physicians regarding the applicability and usefulness of these rules are of paramount importance.

In the future, we intend to develop even more accurate and concise rule extraction algorithms for larger-sized breast cancer datasets and to attempt to come close to achieving rule extraction from Big Data for practical breast cancer screening.

References

[1] Siegel RL, Miller KD, Jemal A. Cancer statistics, 2015. CA Cancer J Clin 2015;65:5–29.

[2] American Cancer Society. Breast cancer facts & figures 2013–2014. Atlanta, Georgia: American Cancer Society, Inc.; 2013.

[3] Rizzo JA, Sherman WE, Arciero CA. Racial disparity in survival from early breast cancer in the department of defense healthcare system. J Surg Oncol 2015;111:819–23.

[4] Subashini T, Ramalingam V, Palanivel S. Breast mass classification based on cytological patterns using RBFNN and SVM. Expert Syst Appl 2009;36:5284–90.

[5] Wolberg WH, Street WN, Mangasarian OL. Image analysis and machine learning applied to breast cancer diagnosis and prognosis. Anal Quant Cytol Histol 1995;17:77–87.

[6] Setiono R. A penalty-function approach for pruning feedforward neural networks. Neural Comp 1997;9:185–204.

[7] Setiono R. Extracting rules from pruned neural networks for breast cancer diagnosis. Artif Intell Med 1996;8:37–51.

[8] Manfrin E, Mariotto R, Remo A, Reghellin AD, Dalfior D, Falsirollo F, Bonetti F. Is there still a role for fine-needle aspiration cytology in breast cancer screening? Cancer 2008;114:74–82.

[9] Nauck D, Kruse R. Obtaining interpretable fuzzy classification rules from medical data. Artif Intell Med 1999;16:149–69.

[10] Salzberg SL. On comparing classifiers: pitfalls to avoid and a recommended approach. Data Min Knowl Discov 1997;1:317–28.

[11] Peňa-Reyes CA, Sipper MA. A fuzzy-genetic approach to breast cancer diagnosis. Artif Intell Med 1999;17:131–55.

[12] Zheng B, Yoon SW, Lam SS. Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. Expert Syst Appl 2014;41:1476–82.

[13] Falco ID. Differential evolution for automatic rule extraction from medical databases. Appl Soft Comput 2013;13:1265–83.

[14] Ǘbeyli ED. Adaptive neuro-fuzzy inference systems for automatic detection of breast cancer. J Med Syst 2009;33:353–8.

[15] Setiono R, Liu H. NeuroLinear: from neural networks to oblique decision rules. Neurocomputing. 1997;17:1–24.

[16] Setiono R. Generating concise and accurate classification rules for breast cancer diagnosis. Artif Intell Med 2000;18:205–19.

[17] Duch W, Adamczak R, Grąbczewski K. A new methodology of extraction, optimization and application of crisp and fuzzy logic rules. IEEE Trans Neural Netw 2001;12:277–306.

[18] Odajima K, Hayashi Y, Tianxia G, Setiono R. Greedy rule generation from discrete data and its use in neural network rule extraction. Neural Netw 2008;21:1020–8.

[19] Karabatak M, Ince MC. An expert system for detection of breast cancer based on association rules and neural network. Expert Syst Appl 2009;36:3465–9.

[20] Hung ML, Hung YH, Cen WY. Neural network classifier with entropy based feature selection on breast cancer diagnosis. J Med Syst 2010;34:865–73.

[21] Marcano-Cedeño A, Quintanilla-Domíngueza J, Andina D. WBCD breast cancer database classification applying artificial metaplasticity neural network. Expert Syst Appl 2011;38:9573–9.

[22] Bhardwaj A, Tiwari A. Breast cancer diagnosis using genetically optimized neural network model. Expert Syst Appl 2015;42:4611–20.

[23] Huysmans J, Setiono R, Baesens B, Vanthienen J. Minerva: Sequential covering for rule extraction. IEEE Trans Syst Man Cybern Part B: Cybern 2008;38:299–309.

[24] Bologna G, Hayashi Y. QSVM: a support vector machine for rule extraction, IWANN Part II, LNCS, vol. 9095. Mallorca, Spain; 2015. p. 276–89.

[25] Bashir S, Qamar U, Khan FH. Heterogeneous classifiers fusion for dynamic breast cancer diagnosis using weighted vote based ensemble. Qual Quant 2015;49:2061–76.

[26] Polat K, Gunes S. Breast cancer diagnosis using least square support vector machine. Digit Signal Process 2007;17:694–701.

[27] Akay MF. Chen Support vector machines combined with feature selection for breast cancer diagnosis. Expert Syst Appl 2009;36:3240–7.

[28] Chen HL, Yang B, Liu J, Liu DY. A support vector machine classifier with rough set based feature selection for breast cancer diagnosis. Expert Syst Appl 2011;38:9014–22.

[29] Chen HL, Yang B, Wang G, Wang SJ, Liu J, Liu DY. Support vector machine based diagnostic system for breast cancer using swarm intelligence. J Med Syst 2012;36:2505–19.

[30] Chen YH, Su CT, Yang T. Rule extraction from support vector machines by genetic algorithms. Neural Comput Appl 2013;23:729–39.

[31] Maldonado S, Weber R. A wrapper method for feature selection using support vector machines. Inf Sci 2009;179:2208–17.

[32] Ster B, Dobnnikar A. Neural networks in medical diagnosis: comparison with other methods. In: Proceedings of the international conference EANN; 1996. p. 427–30.

[33] Abonyi J, Szefert F. Supervised fuzzy clustering for the identification of fuzzy classifiers. Pattern Recognit Lett 2013;24:2195–207.

[34] Goodman D, Boggess L, Watkins A. Artificial immune system classification of multiple-class problems. Proc Artif Neural Netw Eng 2002:179–83.

[35] Şahan S, Polat K, Kodaz H, Güneş S. A new hybrid method based on fuzzy-artificial immune system and k-nn algorithm for breast cancer diagnosis. Comput Biol Med 2007;37:415–23.

[36] Koklu M, Kahramanli H, Allahverdi N. A new approach to classification rule extraction program by the real value coding. Int J Innov Comput Inf Control 2012;8:6303–15.

[37] Fan CY, Chang PC, Lin JJ. A hybrid model combining case-based reasoning and fuzzy decision tree for medical data classification. Appl Soft Comput 2011;11:632–44.

[38] Ǘbeyli ED. A mixture of experts network structure for breast cancer diagnosis. J Med Syst 2005;29:569–79.

[39] Marcono-Cedeňo A, Quintanilla-Domínguez J, Andina D. Breast cancer classification applying artificial metaplasticity algorithm. Neurocomputing 2011;74:1243–50.

[40] Onan A. A fuzzy-rough nearest neighbor classifier combined with consistency-based subset evaluation and instance selection for automated diagnosis of breast cancer. Expert Syst Appl 2015;42:6844–52.

[41] Hassan MR, Hossain MM, Begg RK, Ramamohanarao K, Morsi Y. Breast-cancer identification using HMM-fuzzy approach. Comput Biol Med 2010;40:240–51.

[42] Jaganathan P, Kuppuchamy R. A threshold fuzzy entropy based feature selection for medical database classification. Comput Biol Med 2013;43:2222–9.

[43] Setiono R, Liu H. Symbolic representation of neural networks. IEEE Comput 1996;29:71–7.

[44] Setiono R, Baesens B, Mues C. Recursive neural network rule extraction for data with mixed attributes. IEEE Trans Neural Netw 2008;19:299–307.

[45] Hayashi Y, Tanaka Y, Takagi T, Saito T, Iiduka H, Kikuchi H, et al. Recursive-Rule Extraction Algorithm with J48graft and applications to generating credit scores. J Artif Intell Soft Comput Res (JAISCR) 2016;6:35–44.

[46] 〈http://fiji.sc/javadoc/weka/classifiers/trees/J48graft.html〉; 2016 [last accssed the site 30.09.15].

[47] Webb GI. Decision tree grafting from the all-tests-but-one partition. In: Proceedings of the 16th international joint conference on artificial intelligence (IJCAI), vol. 2; 1999. p. 702–7.

[48] Fortuny EJD, Martens D. Active learning-based pedagogical rule extraction. IEEE Trans Neural Netw Learn Syst 2015;26:2664–77.

[49] University of California. Irvine learning repository, 〈http://archive/ics.uci/edu/m/〉; 2015 [Last accessed 10.01.15].

[50] Witten IH, Frank E. Data mining: practical machine learning tools with Java implementations. San Mateo, California: Morgan Kaufmann, Inc.; 1999.

[51] Quinlan JR. C4.5: programs for machine learning. Morgan Kaufmann series in machine learning. San Mateo, California: Morgan Kaufman, Inc.; 1993.

[52] Quinlan JR. Induction of decision trees. Mach Learn 1986;1:81–106.

[53] Webb GI. Decision tree grafting, leairning. In: Proceedings of the 15th international conference on artificial intelligence (IJCAI), IJCAI'97, vol. 2; 1997. p. 846–85.

[54] Nagar S, Iacco A, Riggs T, Kestenberg W, Keidan R. An analysis of fine needle aspiration versus cone needle biopsy in clinically palpable breast lesions: a report on the predictive values and cost comparison. Am J Surg 2012;204:193–8.

[55] 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.

[56] Isa NSM, Subramaniam E, Mashor MY, Othman NH. Fine needle aspiration cytology evaluation for classifying breast cancer using artificial neural network. Am J Appl Sci 2007;4:999–1008.

[57] McCluggage WG, McManus DI, Caughley LM. Fine needle aspiration (FNA) cytology of adenoid cystic carcinoma and adenomyoepithelioma of breast: two lesions rich in myoepithelial cells. Cytopathology 1997;8:31–9.

[58] Wieczorek TJ, Krane JF, Domanski HA, Åkerman M, Carlén B, Misdraji J, et al. Cytologic findings in granular cell tumors, with emphasis on the diagnosis of malignant granular cell tumor by fine-needle aspiration biopsy. Cancer 2001;93:398–408.