Informatics in Medicine Unlocked 1 (2015) 9–16

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.