Computer Methods and Programs in Biomedicine 153 (2018) 259–268

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

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

Robust breast cancer prediction system based on rough set theory at National Cancer Institute of Egypt

Saeed Khodary M. Hamouda[a,*], Mohammed E. Wahed[b], Reda. H. Abo Alez[c], Khaled Riad[d]

[a] *Computer Laboratories, Zagazig University, Zagazig 44519, Egypt*
[b] *Faculty Of Computers and Informatics, Suez Canal University, Ismailia, Egypt*
[c] *Faculty of Engineering, Al-Azhar University, Cairo, Egypt*
[d] *Mathematics Department, Faculty of Science, Zagazig University, Zagazig 44519, Egypt*

ARTICLE INFO

Article history:
Received 3 January 2017
Revised 10 September 2017
Accepted 12 October 2017

Keywords:
Breast cancer
Prediction and diagnosis
Rough set theory
Approximation
Reduction
MATLAB

ABSTRACT

Background: Breast cancer is one of the major death causing diseases of the women in the world. Every year more than million women are diagnosed with breast cancer more than half of them will die because of inaccuracies and delays in diagnosis of the disease. High accuracy in cancer prediction is important to improve the treatment quality and the survivability rate of patients.

Objectives: In this paper, we are going to propose a new and robust breast cancer prediction and diagnosis system based on the Rough Set (RS). Also, introducing the robust classification process based on some new and most effective attributes. Comparing and evaluating the performance of our proposed approach with the clinical, Radial Basis Function, and Artificial Neural Networks classification schemes.

*Methods***:** The dataset used in our experiments consists of 60 samples obtained from the National Cancer Institute (NCI) of Egypt. We have used the RS theory to robustly find dependence relationships among data, and evaluate the importance of attributes through:

- Applying the Approximation Sets on this dataset to identify the patient's cancer stage (0, IA, IB, IIA, IIB, IIIA, IIIB, IIIC, IV); and
- Running the Reduction process on this dataset to identify which attributes (symptoms) are most effective for description and predict breast cancer stage.

Results:

- Our approach has classified the patients into 9 different stages, Stage 0 with accuracy 75%, Stage IA with accuracy 71%, Stage IB with zero patients, Stage IIA with accuracy 86%, Stage IIB with accuracy 67.5%, Stage IIIA with accuracy 85%, Stage IIIB with accuracy 100%, Stage IIIC with zero patients, and Stage IV with accuracy 100%;
- The Reduction process gives as output, the most effective symptoms to early predict and accurately diagnosis the breast cancer. That are represented in Lymph Node Status, Tumor Size, Estrogen Receptor Status, Progesterone Receptor Status, and Metastasis; and
- The last but not least, we have found two patients (patient No. 11 and 51 from our dataset) in High Risk Status, which requires intensive and special treatment.

Conclusion: We have introduced the robustness of the RS theory in early predicting and diagnosing the breast cancer. This lay more importance to the contribution and efficiency of RS theory in the field of computational biology.

© 2017 Elsevier B.V. All rights reserved.

[*] Corresponding author.
E-mail addresses: hamoudasaeed@yahoo.com (S.K.M. Hamouda), khaled.riad@science.zu.edu.eg (K. Riad).

https://doi.org/10.1016/j.cmpb.2017.10.016
0169-2607/© 2017 Elsevier B.V. All rights reserved.

1. Introduction

The body is made up of trillions of living cells, normal body cells grow and divide into new cells and die in an orderly way. Cancer begins when cells of the body start to divide without stopping (i.e. out of control) [1,2]. Breast cancer is characterized by the uncontrolled growth of abnormal cells in the milk producing glands of the breast, It is the most common neoplasm among women in the majority of the developed countries, accounting for one-third of newly diagnosed malignancies [3]. It is a highly heterogeneous disease, encompassing a number of biologically distinct entities with specific pathologic features and biological behaviors [3,4]. Different breast tumor subtypes have different risk factors, clinical presentation, histopathological features, outcome, and response to systemic therapies. Thus, an accurate and early classification of breast cancer is urgently required. Improvements in prevention and diagnosis have resulted in earlier diagnosis and treatment. Earlier detection of cancer is curable and may increase the survivability, but detecting cancer in earlier stage is difficult [1,2,5,6].

Large amounts of data about the patients with their medical conditions are presented in the medical databases. Breast cancer is one of the most important medical problems. The growth of the amount of data and the number of existing databases far exceeds the ability of humans to analyze this data. Analyzing all these databases is one of the difficult tasks in the medical environment. Thus, there is both a need and an opportunity to extract knowledge from databases. Medical databases have a large quantity of information about patients and their medical conditions. In order to accomplish move through the correct treatment, physicians classify the individual breast cancer according to standard parameters that include type, grade, stage, and gene expression of the breast cancer. In this paper, we are interested in classifying breast cancer according to stage, to describe the correct treatment as early as possible. For screening purposes, mammography is quite often used hence it gives the maximum possibilities for a physician to trace the exact location of micro calcifications and other possible indicators in the breast tissue.

In this paper, we are doing the classification accuracy of the TNM staging process using the RS theory. Also, the results are compared with the previously proposed RS, RBF, and ANN. In this study, the total number of patients with breast cancer studied are 60. For all the types of classifications, the input variables are nothing but the TNM variables (such as Metastasis (M), Tumour Size (TS), Lymph Node Status (LNS), Estrogen Receptor Status (ERS), Progesterone Receptor Status(PRS), Histological Type (HT), and Histological Grade (HG)). To increase the classification accuracy, we have consulted and the physicians Tumor Department at both Zagazig University Hospitals and Suez Canal University Hospitals, Egypt, about the most effective and redundant attributes.

The theory of RS theory is a mathematical tool for extracting knowledge from uncertain and incomplete database information [7,8]. The theory of RS can be used to find dependence relationships among data, and evaluate the importance of attributes. The theory assumes that we first have the necessary information or we have exactly the same information of two objects then we say that they are indiscernible (similar), i.e., we cannot distinguish them with known knowledge.

1.1. Motivation

To the best of our knowledge and according to the most recent literature in breast cancer classification and treatment, most of the proposed approaches are working on international databases such as WBC data set. Also, they are doing the classification process based on some static and old attributes. As well as, there is no clear treatment description. This is behind our motivation to propose a novel intelligent approach for breast cancer prediction and diagnosis based on rough set theory, new most effective attributes, and High Risk Status (HRS). The patient is classified as a HRS, if $Age \leq 35 \wedge (ERS = +ve \vee PRS = +ve) \wedge T > 2cm \wedge HG = 2 : 3$, for ERS and PRS to be $+ve$ means that their values is greater than 50. This is due to the discussion with specialist physicians at Zagazig University Hospitals and Suez Canal University Hospitals.

Also, the authors in [9] have introduced a comprehensive breast cancer classification with Radial Basis Function and Gaussian Mixture Model, which is only based on five classification stages. Thus, we have to introduce nine classification stages, to be more accurate and specific.

1.2. Organization

The rest of this paper is organized as follows: Section 2 provides the related literature and discussion, which introduce a discussion for the breast cancer classification with two well known methods (RBF, and ANN) and other different contributions in that field. Section 3 describes an overview on breast cancer and Rough Set theory basic concepts. Section 4 presents our proposed approach with its materials and methods, approximations and accuracy, and reduction and core attributes. The proposed approach implementation is presented in Section 5. Section 6 provides a detailed comparison for our approach with the clinical, RBF, and ANN classifications. As well as, our obtained results. This is followed by the conclusion Section 7.

2. Related work and disscussion

2.1. Rough set theory and breast cancer

Chul-Heui Lee et al. [10] proposed a new classification method based on the hierarchical granulation structure using the rough set theory. The hierarchical granulation structure was adopted to find the classification rules effectively. The classification rules had minimal attributes and the knowledge reduction was accomplished by using the upper and lower approximations of rough sets. A simulation was performed on WBC dataset to show the effectiveness of the proposed method. The simulation result showed that the proposed classification method generated minimal classification rules and made the analysis of information system easy. On the other hand, the authors did not consider the attribute reduction problem. Another paper [11] also working on the WBC data set using rough set theory. This paper presented a rough set method for generating classification rules from a set of observed 360 samples of the WBC data. The attributes were selected, normalized and then the rough set dependency rules were generated directly from the real value attribute vector. Then the rough set reduction technique was applied to find all reducts of the data which contains the minimal subset of attributes that are associated with a class label for classification. The authors showed that the total number of generated rules was reduced from 472 to 30 rules after applying the proposed simplification algorithm.

It is clear that most of the proposed approaches using rough set theory have no comparison nor interaction with physicians for doing real tests with actual datasets. In our method we have trained our approach with actual data sets from the National Cancer Institute (NCI) of Egypt [12].

2.2. Radial basis function and breast cancer

There integration between the Relief-F algorithm and the Radial Basis Function has been proposed in [13]. The Relief-F algorithm has been chosen as the method of the feature selection. The Radial

Download English Version:

https://daneshyari.com/en/article/6891237

Download Persian Version:

https://daneshyari.com/article/6891237

Daneshyari.com