ISSN: 0067-2904

Diagnosis the Breast Cancer using Bayesian Rough Set Classifier

Ayad R. Abbas *, Marwa A. Shihab
Department of Computer Science, University of Technology, Baghdad, Iraq

Abstract

Breast cancer was one of the most common reasons for death among the women in the world. Limited awareness of the seriousness of this disease, shortage number of specialists in hospitals and waiting the diagnostic for a long period time that might increase the probability of expansion the injury cases. Consequently, various machine learning techniques have been formulated to decrease the time taken of decision making for diagnoses the breast cancer and that might minimize the mortality rate. The proposed system consists of two phases. Firstly, data pre-processing (data cleaning, selection) of the data mining are used in the breast cancer dataset taken from the University of California, Irvine machine learning repository in this stage we modified the Correlation Feature Selection (CFS) with Best First Search (BFS) established on the Discriminant Index (DI) so as to reduce the complexity of time and get high accuracy. Secondly, Bayesian Rough Set (BRS) classifier is applied to predict the breast cancer and help the inexperienced doctors to make decisions without need the direct discussion with the specialist doctors. The result of experiments showed the proposed system give high accuracy with less time of predication the disease.

Keywords: The breast cancer, Correlation Feature Selection (CFS), Bayesian Rough Set (BRS), Discriminant Index (DI).

تشخيص سرطان الثدي باستخدام نظرية التصنيف *Bayesian Rough Set*

أياد روضان عباس *، مروة احمد شهاب
قسم علوم الحاسوب، الجامعة التكنولوجية ، بغداد ، العراق

الخلاصة

ان سرطان الثدي واحد من الاسباب الاكثر شيوعا للموت بين النساء في العالم. ان قلة الوعي بخطورة هذا المرض وقلة عدد الاختصاصين في هذا المجال،كذلك الانتظارلوقت طويل لأجل الحصول على نتائج التشخيص ادى لزيادة حالات الاصابة. نتيجة لذلك، استخدمت العديد من تقنيات تعليم الماكنة لتقليل الوقت اللازم لاتخاذ القرار، ذلك قد يساهم تقليل حالات الوفاة. هذا البحث يقدم نظام مقترح والذي يتكون من مرحلتين اعتمادا على بيانات التي تم الحصول عليها من مستودع تعليم الالة جامعة كاليفورنيا في ايرفين. اولا معالجة البيانات (تنظيف البيانات واختيار الخصائص التي تؤثر على اتخاذ القرار) وفي هذة المرحلة تم تطوير نموذج اختيار الخصائص المرتبطة CFS مع استخدام خوارزمية البحث الاستدلالي BFS استنادا الى مؤشر التمايز DI من اجل تقليل الوقت وزيادة الدقة بالتشخيص. ثانيا استخدام طريقة BRS

*Email:rashed2221@yahoo.com

وهي طريقة مطورة استخدمت لتشخيص سرطان الثدي ومساعدة الأطباء المبتدئين في اتخاذ القرار دون الحاجة الى التشاور المباشر مع الاخصائيين. اظهرت نتائج التجريبية ان النظام المقترح يعطي دقة عالية بأقل وقت ممكن لتشخيص المرض.

Introduction

The breast cancer is the most common types of cancer among women in all over the world. It is assessed that 1 in 8 women alive today in the United States of America will be diagnosed with breast cancer during her lifetime. An expected 232,670 women will be diagnosed with and 40,000 women will die of cancer of the breast in 2014 [1]. The proposed system consists of two phases. First, data cleaning is applied to remove noise data, and remove the missing value in the data. Then, the Correlation Feature Selection (CFS) with Best First Search (BFS) are combined and modified to explored the breast cancer disease dataset taken from the University of California, Irvine machine learning repository [2], with the purpose of effectively identifying the several attributes, which best predict a selected target attribute. Second, Bayesian Rough Set (BRS) classifier is applied to significantly predict the breast cancer mortality.

Related Works

There are many researches applied on the breast cancer diagnosis with Wisconsin Breast Cancer Database (WBCD) and most of them have high accuracy, these researches are listed as follows:

1. Support vector machine and neural network: [3] has been implemented different machine learning techniques (SVM, k-mean and ANN), the results show SVM is an effective and accurate method for the breast cancer diagnosis, [4] medical decision system has been proposed using an SVM algorithm for benign/malignant the breast cancer classification, [5] has explored the applying SVM and compare with Bayesian classifier and ANN for prognosis and diagnosis the breast cancer disease**,** [6] founded SVM suited for the diagnosis when compared with (KNN ,naïve Bayesian) classifiers .
2. Fuzzy set: [7] the medical diagnosis problem of the breast cancer is solved effectively by using a fuzzy genetic approach, [8] a method was obtained by using hybridizing fuzzy artificial immune system with K-nearest neighbour algorithm to solve the breast cancer diagnosis problem.
3. Neural network: [9] the performance of statistical neural network structure ,redial basis network (RBF),general regression neural network(GRNN) and probabilistic neural network (PNN) are examined on the breast cancer dataset to increase the accuracy and objectivity of the diagnosis, [10]association rules and neural network (AR+NN) model are presented for detecting the breast cancer disease and obtain fast automatic diagnosis system,. [11] investigate the performance of neural network with adaptive resonance theory (ART) structure for the breast cancer diagnosis problem.
4. Decsion tree: [12] has explored the applying range of techniques (random tree, Quinlan's C4.5 decision tree) looking for the best performance in the breast cancer diagnosis, [13] analyses the performance different machine learning algorithms (viz, random tree, ID3, ART, C4.5, naïve Bayesian) to serve the best accuracy for diagnosis, [14] investigated the implementation of different classification techniques (SMO, IBK, BF tree) for diagnostic problem.
5. Rough set: [15] present a rough set method for generating classification from set of the breast cancer data, [16] rough set based on supporting vector machine classification (RS-SVM) is proposed for the breast cancer diagnosis. 6-Least square support vector machine: [17] the effectiveness of LS-SVM is evaluated onset of the breast cancer data and the proposed system obtain very promising accurate decision in classifying the breast cancer patients.

The Proposed Materials and Methods

In this section, the proposed system applies different data mining techniques on the breast cancer data set and beginning with a training set on the breast cancer patient's dataset: data pre-processing, Features Selection Algorithm (CFS with BFS) and machine learning algorithm.

The Breast Cancer Diseases Dataset [2]

In this paper, the University of California, Irvine (UCI) data sets of the breast cancer are applied as a part of the research. The UCI The breast cancer data sets of 699 patients are collected from the university of Wisconsin hospitals, Madison from William H. Walberg. The 9 attributes information for these data sets are related to fine needle aspirates taken from human the breast cancer tissue, each of

these attributes represents as an integer value between 1 to 10 with two classes 2 for benign and 4 for malignant as shown in Table- 1.

Table 1- Complete Breast Cancer Diseases Dataset Description [2].

Label	Attribute
t1	Thickness of clusters
t2	Homogeneity of Cell Size
t3	Homogeneity of Cell Shape
t4	Marginal Adhesion
t5	Single Epithelial Cell Size
t6	Bare Nuclei
t7	Bland Chromatin
t8	Normal Nucleoli
t9	Mitoses

Proposed Data Mining Model

In this section, the proposed intelligent model is shown in Figure-1. This model consists of two steps: preprocessed data and machine learning tasks. The preprocessed data consist of two sub steps: data cleaning and the feature selection. The features selection task is proposed using CFS, rough set and Best first search.

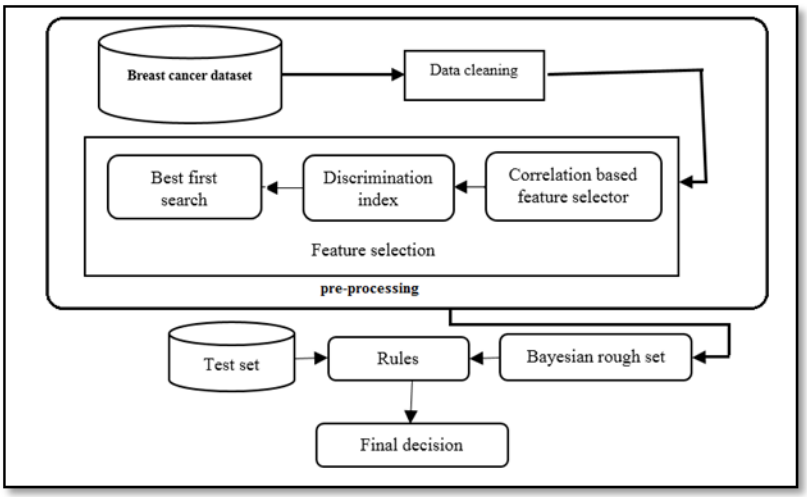

Figure 1- Proposed Model

Preprocessed of the Breast Cancer Dataset

Data Cleaning: Before the BRS process and to enhance accuracy of data analysis and classification, data cleaning pre-processing has been applied. The data cleaning task is applied for removing the patient data that content missing values.

Correlation Based Feature Selector (CFS) [18]

In this section, CFS is used to select relevant attributes. CFS is a one of the simple selector algorithms that using an evaluation function for sorting feature subsets considers the degree of correlation. The effect of the heuristic evaluation function is to subsets that have attributes that are associated extremely with the class and unassociated with each other. In this algorithm, the attributes that unrelated to class (low merit) and related to other attributes (high merit) that will be reduced from the information table. CFS's feature subset heuristic function represented in (Equation 1):

$$\mathbf{Merit}_s = \frac{k\,\overline{rcf}}{\sqrt{+k(k-1)\overline{rff}}} \quad (1)$$

Where M_S is the merit of a subset attributed S with k attributes, r_{cf} is the correlation degrees between the attributes and the class, and r_{ii} is the inter correlation degree between attributes using DI.

DI is utilized to calculate the correlations between the features and the class Because of all the breast cancer dataset features are nominal, where DI denoted the ratio of the number of positive instances to the number of all instances using Rough Set Methods in (Equation 2).

$$DI = 1 - \frac{The\ set\ all\ boundary\ region}{The\ set\ of\ all\ training\ examples} \quad (2)$$

Best first search (BFS) Method [19]

Best first search is one of the search artificial intelligence methods, the search path working with backtracking. It uses the search space for moving the local search to the current attribute subset. If the path being examined starts to look hopeless, the BFS can backtrack to another hopeless past subset and continue the search from there. At the time, a BFS examined the whole search space to apply the stopping condition normally as shown in the following algorithm:

Algorithm 1-BFS Algorithm

OPEN_list = [start state]
CLOSED-list = empty_list
While OPEN_list is not empty
Do

1. *Remove the best node from OPEN-list, call it n that have high merit, and add it to CLOSED-list.*
2. *If n is the goal state, backtrack e path to n (through recorded parents) and return path.*
3. *Create n's child.*
4. *for each child do:*

a. If it is not in CLOSED_list and it is not in the OPEN_list: add it to OPEN-list after evaluating it, and record its parent.
b. Otherwise, if this new path is better than the previous one, change its recorded parent.
I. If it is not in the OPEN _ list, add it to the OPEN _list.
II. Otherwise, change its priority in the OPEN _ list with the new evaluation.

Done.

Implementation Using Bayesian Rough Set (BRS) [20]

Firstly, starting with an original RS data analysis are based on the dataset, called an information system (IS). An information system is a data content in the table, whose columns are categorized by attributes, rows are categorized by interest objects and entries of the table are attribute values. IS is a pair of (U, A), where U symbolize non-empty finite set of objects, A symbolize non-empty finite set of attribute such T is an element of A is identified function a: u → VT , with ergard to VT symbolize the values set of a. Any B⊆ A establishes a bilateral relation I(B) on U, which will be called an indiscernibility relation the partition space U/B are obtained, called the B-indiscernibility relation INDs(B),where elements e is an element of U/B are called the B-indiscernibility classes of objects.

In Table-2, decision system S = (U, A ∪ {class}), U = {U1,…, U683}, and A = {t1,…, t9}

As in Figure-1, the BRS purposes to provide general rules with three main processes; extract appropriate the rules of decision from a decision that content in the table: the knowledge decrease, which is a procedure associated with ignoring unnecessary condition features from the table of decision. Or its not every condition feature is necessary for classify the objects in the (IS).

The approximation regions can represent by using the difference between subsequent probability P (X|E) and previous probability P (X) is as follows:

$$\mathbf{d'(X|E) = P(X|E) - P(X)} \quad \text{.....(3)}$$

By using this relevance measure we can divide the data into positive, negative and boundary regions

$$\begin{aligned} \mathbf{POS^{\alpha}(X_t)} &= \mathbf{d(X_t,E_i) >= \alpha_t} \\ \mathbf{NEG^{\alpha}(X_t)} &= \mathbf{d(X_t,E_i) <= -\alpha_t} \\ \mathbf{BND^{\alpha}(X_t)} &= -\alpha_t < d(X_t,E_i) < \alpha_t \end{aligned} \quad \text{..... (4)}$$

$$\alpha_t = \min\left(P(X_t), \sum P(X_c)\right) \quad \text{..... (5)}$$

where:

α_t: The single parameter that take a great role in controlling the significant approximation degree.

In the extraction process of decision rules, Discriminant index η is applied to provided a way for computing the certainty degree in classifying the objects set, in which the greatest index value establishes the most effective attribute:

$$\eta = \frac{card(BND(X_t))}{card(U)} \quad \text{..... (6)}$$

Finally, a BRS-based recommender agent was proposed to a group of patients to guide them in their treatment, and specify the areas they must concentrate on according to the way they fit the decision rules.

Experimental Results and Evaluation

This section presents the experimental results and evaluation in discovering main factors for the breast cancer. The breast cancer data set is eliminated in identical records by removing the missing value. After that, selected relevant breast cancer attributes by applying CFS. The correlation scores of the breast cancer dataset are calculated by using rough set method.

The forward best first search uses the search space for feature selection by applying Equation (1) to calculation the merit for each feature depending on the higher correlation between particular attributes and class, and the lower interior correlation between attributes. The higher correlation degrees of an attributes t1, t7, t2, t3, t5, t8, t4 and t6 with class are 0.796, 0.546, 0.453, 0.493, 0.389, 0.367, 0.424 and 0.411 respectively. The BFS starts with the empty set of attributes (zero merit). When K=1 means the evaluation of each single attribute added to the empty set; t1 has the highest score so is added to the subset. The next step when K=2 includes trying each of the remaining features with t1 and choosing the best (t7). Because the score of attribute (t7) with class is 0.546 (high score) and 0.37 (low score) with the t1 attributes. As a results the relevant features selection are t1, t7, t2, t3, t5, t8, t4 and t6 are shown in Table- 2.

Table 2- The breast cancer Dataset after Pre-processing

Domain	**t**1	**t**2	**t**3	**t**4	**t**5	**t**6	**t**7	**t**8	Class
U1	4	1	1	1	2	3	1	1	2
U2	10	5	6	10	6	10	7	7	4
U3	5	10	10	3	8	1	5	10	4
U4	3	1	1	1	2	1	3	2	2
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
U683	4	5	10	10	10	2	2	10	4

The correctly classified instances are 682 with 99.85% and incorrectly a classified instance is 1 with 0.146%.

Experimental results reveal that the proposed system has excellent accuracy. Table-3 shows the differences between the proposed method and other machine learning algorithms (ID3 Decision Tree, J.48 Decision Tree and Bayesian Network) using the same dataset, which was conducted with and without features selection. However, Table-4 shows the performance for training data classification which can be compared with other researches.

Table 3-Related Classifier Methods

Classifier	**Correctly Classified Instances**	**Features selection**	**Execution Time**
Proposed BRS	99.85 %	yes	0.00004 s
Proposed BRS	99.0 %	no	0.00001s
ID3	91.65 %	yes	0.02 s
J.48	93.41 %	yes	0.02 s
Net Bayes	97. 80 %	yes	0.02 s
NaiveBayes	95.99%	no	0.01s
J.48	92.56%	no	0.09s
Net Bayes	97.13%	no	0.04s
ID3	90.41%	no	0.03s

Table 4-Performance for Training Data Classification from Literature

Authors	Year	Machine Learning Tools	Accuracy %
Tüba Kıyan, et al. [9]	2003	Artificial Neural Network	96.40
Seral Şahan, et al. [8]	2007	fuzzy-artificial immune system and K-NN	98.14
Hui-Ling Chen, et al. [16]	2011	SVM+Rough set	96.00
Mandeep Rana, Pooja Chandorkar [6]	2015	SVM-linear	63.64
Mandeep Rana, Pooja Chandorkar [6]	2015	Logistic regression-generalized	92.59

Conclusion

In this research many methods have been used to diagnose the breast cancer on University of California, Irvine dataset. The modification of Correlation Based Feature Selector (CFS) model using Discrimination Index (DI) and Best First Search (BFS) in addition to the Bayesian Rough Set (BRS) are used to develop the prediction and classification model. The proposed system has been examined to provide 99.85 % classification with high accuracy and less time of disease predication in order to assist the inexperienced doctors for their final decision making on their patient without need the direct discussion with specialists.

References

1. Siegel, R., Ma, J., Zou, Z. and Jemal, A. **2014**. Cancer statistics. *Cancer Journal for Clinicians*, **64** (1): 9-29.
2. William, H. Wolberg. **1991**. University of Wisconsin Hospitals Madison, Wisconsin, USA, Dataset available: http://www.archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/.
3. Liu, Hx., Zhang Rs., Luan, F., Yao, Xj., Liu, Mc., Hu, Zd. and Fan, Bt. **2003**. Diagnosing the breast cancer based on support vector machines. *Journal of Chemical Information and Computer Sciences*, **43**(3): 900-907.
4. Akay, M.F. **2009**. Support vector machines combined with feature selection for the breast cancer diagnosis. *Journal of Expert Systems with Applications*, **36**(2): 3240–3247.
5. Maglogiannis, I., Zafiropoulos, E., Anagnostopoulos, I. **2009**. An intelligent system for automated the breast cancer diagnosis and prognosis using SVM based classifier. *The International Journal of Artificial Intelligence*, **30**(1): 24-36.
6. Mandeep R., Pooja C., Alishiba D., Nikahat K. **2015**. Breast cancer diagnosis and recurrence prediction using machine learning techniques. *International Journal of Research in Engineering and Technology*, **4**(4): 372-376.
7. Pena-Reyes, C. and Sipper, M. **1999**. A fuzzy-genetic approach to the breast cancer diagnosis. *Journal of Artificial Intelligence in Medicine*, **17**(2), pp: 131-155.
8. Sahan, S., Polat, K., Kodaz, H., Güneş, S. **2007.** A new hybrid method based on fuzzy-artificial immune system and k-NN algorithm for breast cancer diagnosis. *Journal of Computers in Biology and Medicine*, **37**(3):415-423.
9. Tüba K., Tülay Y. **2003**. Breast cancer diagnosis using statistical neural networks, International XII. Turkish Symposium on Artificial Intelligence and Neural Networks, January, pp:1149-1153, Turkey.
10. Murat, K., Cevdet I. **2009.** An expert system for detection of the breast cancer based on association rules and neural network. *Journal of Expert Systems with Applications*, **36**(2): 3465-3469.
11. Sonia, N., Verma, H. and Uday, S. **2012**. A Review of the breast cancer detection using ART model of neural networks. *International Journal of Advanced Research in Computer Science and Software Engineering*, **2**(10): 311-318.

12. Jacob, S., Ramani, G. **2012.** Efficient classifier for classification of prognostic the breast cancer data through data mining techniques. Proceedings of the World Congress on Engineering and Computer Science 2012, October 24-26. San Francisco, USA.
13. Shajahaan, S., Shanthi, S. and ManoChitra, V. **2015**. Application of data mining techniques to model the breast cancer data. *International Journal of Emerging Technology and Advanced Engineering*, **3**(11): 362-369.
14. Vikas, C. and Saurabh, P. **2014.** A novel approach for the breast cancer detection using data mining techniques. *International Journal of Innovative Research in Computer and Communication Engineering*, **2**(1): 2456- 2465.
15. Aboul, Ella Hassanien and Jafar, M.H. Ali. **2004**. Rough set approach for generation of classification rules of the breast cancer data, *Journal of Informatica.* **15**(1): 23-38.
16. Chen, H., Yang, B., Liu, J., Liu, D. **2011**. A support vector machine classifier with rough set-based feature selection for the breast cancer diagnosis. *Journal of Expert Systems with Applications*, **38**(7): 9014-9022.
17. Kemal, Polat , Salih, Güne. **2007**. The breast cancer diagnosis using least square support vector machine. *Journal of Digital Signal Processing*, **17**(1): 694–701.
18. Mark, A. Hall, **1999**. Correlation-based Feature Selection for Machine Learning. PhD. Thesis, Department of Computer Science, the University of Waikato, Hamilton, New Zealand.
19. Rich, E. and Knight, K. **2009**. *Artificial Intelligence*. Third edition, McGraw-Hill.
20. Ayad, R. Abbas, Liu. Juan and Safaa, O. Mahdi. **2007.** A New Version of the Bayesian Rough set based on Bayesian Confirmation Measure, International Conference on Convergence Information Technology, November pp: 284-289, China.