International Journal of Medical, Medicine and Health Sciences
ISSN: 2517-9969
Vol:5, No:5, 2011

# Feature Selection for Breast Cancer Diagnosis: A Case-Based Wrapper Approach

Mohammad Darzi, Ali AsgharLiaei,Mahdi Hosseini, HabibollahAsghari

***Abstract*—**This article addresses feature selection for breast cancer diagnosis. The present process contains a wrapper approach based on Genetic Algorithm (GA) and case-based reasoning (CBR). GA is used for searching the problem space to find all of the possible subsets of features and CBR is employed to estimate the evaluation result of each subset. The results of experiment show that the proposed model is comparable to the other models on Wisconsin breast cancer (WDBC) dataset.

***Keywords*—**Case-based reasoning; Breast cancer diagnosis; Genetic algorithm; Wrapper feature selection

## I. Introduction

BREAST cancer is one of the most common cancers among women World Wide and its incidence is about one million new patients annually by the year 2000. There is an overall increase of 2% in the incidence of breast cancer throughout the world per year. Worldwide it is estimated that 420000 deaths would occur annually as a result of breast cancer by the year 2000. Although breast cancer is a potentially fatal condition, early diagnosis of disease can lead to successful treatment [1]. One of the important steps to diagnose the breast cancer is classification of tumor. Tumors can be either benign or malignant but only the latter is cancer. So, malignant tumors generally are more serious than benign tumors. Early diagnosis needs a precise and reliable diagnosis procedure that allows physicians to distinguish between benign breast tumors and malignant ones [2]. For this purpose, there are various computer-based solutions to serve as the diagnosis procedure and assist the physicians to specify the type of breast mass. These systems, called Medical Diagnostic Decision Support (MDDS) systems, can augment the natural capabilities of human diagnosticians incorporating imprecise models about the incompletely understood and exceptionally complex process of medical diagnosis [3].

One of the problems in these systems is the multiplicity of features. Many of these features may be irrelevant to the mining task or redundant [4]. Therefor these features increase the cost of retain and management of data and cause of confusing the algorithm of classification. Generally, they lead to a low learning precision [5, 6, 7]. It can be proposed some methods that can cope with this problem. One of them is feature selection [8]. Feature selection task is to choose a subset of the original features present in a given dataset that provides most of the useful information [9]. Feature selection has many advantages; some benefits include facilitating data visualization and data understanding, reducing the measurement and storage requirements, reducing training and utilization times, defying the curse of dimensionality to improve prediction performance [10].

There are three approaches for feature selection: Wrapper, Filter and Embedded [10]. In wrapper approach, the selected subset of features is evaluated by a machine learning algorithm that is the classification engine. Filter approach uses some techniques to score the selected subset, ignoring classifier algorithm. In embedded approach, selecting the best subset of features is performed during the process of training.

The filter approach has a drawback. In this approach the process of selecting the best subset of features is independent to the classifier engine. It might cause a bad effect on the output of classifier algorithms because the subset is just selected based on correlation between data. The wrapper and embedded approaches don't have the mentioned drawback because wrapper uses the same method for evaluating the selected subset of features that is used for classification and embedded approach performs feature selection during the process of training and it is not independent of the classifier algorithm. By using the learning machine as a black box, wrappers are remarkably universal and simple [10].

The first step of wrapper based feature selection methods is search among the wide variety of possible subsets of features. A search algorithm can be employed to perform this step. In many studies GA is used [11, 12, 13, 14, 15, 16].

GA invented by John Holland in the 1960s and developed by him and his students and colleagues at the University of Michigan in the 1960s and the 1970s [17]. Holland in [18] presented GA as an abstraction of biological evolution and gave a theoretical framework for adaptation under GA. GA is a search algorithm that models the natural process biological evolution. For every problem, there is a solution space that genetic try to find the optimal solution for the specific problem by using some operators such as mutation, crossover, selection and etc. In feature selection problem, the optimal solution is a subset of features which has the best result. Each subset of features represents as chromosome in GA.

In wrapper feature selection approach the algorithm that is used as classifier should be employed to evaluate each subset which is selected by GA. In this study CBR is used. CBR is a methodology that provides the ability to use past experiences

M. Darzi is Faculty Member of ICT Research Institute-ACECR (No. 5,Saeedi Alley, Kalej Intersection, Enghelab Ave., Ferdowsi Sq., Tehran, Iran, E-mail: Modarzi@ictrc.ir).

A. Liaei is Member of Information System Research Group, ICT Research Institute-ACECR,E-mail: Liaei@ictrc.ir).

M. Hosseini is Member ofInformation System Research Group, ICT Research Institute-ACECR, E-mail: Hosseini@ictrc.ir).

H. Asghari is General Manager of ICT Research Institute-ACECR, E-mail: Asghari@ictrc.ir).

to solve new problems. In CBR approach, problems are solved by adapting solution of prior problems to new problem's context. The four fundamental steps of CBR are [19] (Fig. 1):

- Retrieve some cases based on a similarity measure;
- Reuse the selected cases to solve a given problem;
- Revise the proposed solution if needed, based on the fact that the new problem and matched case partially differ;
- Retain the problem and its solution as a pair in case base.

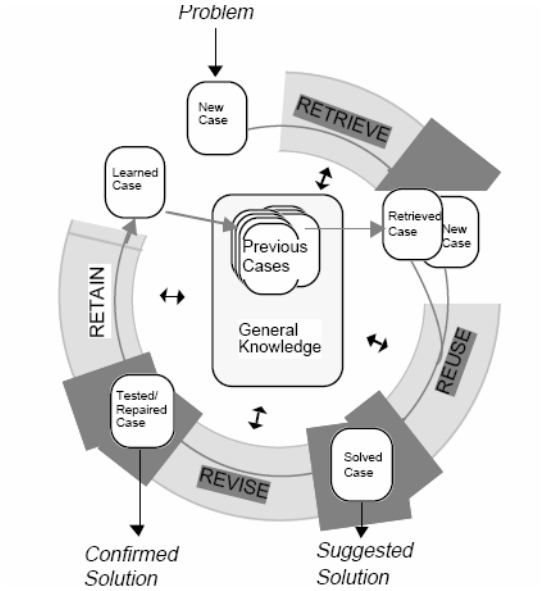


Fig.1 CBR Cycle

In this paper, a wrapper feature selection model for breast cancer diagnosis is proposed. This model employed GA for search phase and CBR for evaluation phase. The model evaluated on Wisconsin Diagnostic Breast Cancer (WDBC) Data.

There are some related works but in these studies there are some differences between them and proposed model. For example, Ahn et al in [20,21] introduced a feature selection using CBR and GA But they used feature and instance selection simultaneously. This method is called Simultaneous optimization approaches [20].The main idea of the simultaneous optimization approach is to reduce the dimensions of features and instances simultaneously. Hence, the chromosome representation of this model is different from the proposed model. The second difference is the domain. The domain of mentioned model is customer classification whereas the domain of proposed model is breast cancer. Also, Ahn et al in [22] proposed a model as same as [21]. But in [22] they used a commercial domain in their study. KYOUNG-JAE KIM in [23] used a case-based feature selection using genetic algorithm in financial domain. Beddoe and Petrovic in [24] employed CBR and GA for both feature selection and feature weighting simultaneously. So, the chromosome representation is different to the proposed model. Also their domain is different too. In another study, Jarmulak et al in [25] used GA for case-based feature selection and feature weighting in drug domain. Golobardes et al in [26] used a feature selection and instance selection simultaneously based on CBR and GA for two dimensional reductions of mammogram images.

## II. Architecture Of Model

The aim of proposed model is selecting best subset of features that is caused the classifier model to have the optimal performance. For this reason, a case-based wrapper feature selection with GAis designed. In this model GA is used for search all possible subsets of features and CBRis employed for evaluating each subset.

As the following, the process of feature selection is presented.

### *A. Chromosome representation*

In proposed model, each chromosome is a subset of features. The size of chromosome (number of genes) is equal to the number of features that represent the specification of a cancer patient. A chromosome is represented in form of binary string that is 0 or 1. 1 means the corresponding feature is selected and 0 means it is not selected (Fig. 2). As shown in figure, n is the number of gene in chromosome (size of chromosome).

### *B. Population*

A population is a set of chromosomes. In proposed model, the first population is generated randomly. The number of chromosome in each population (size of population) is 100.

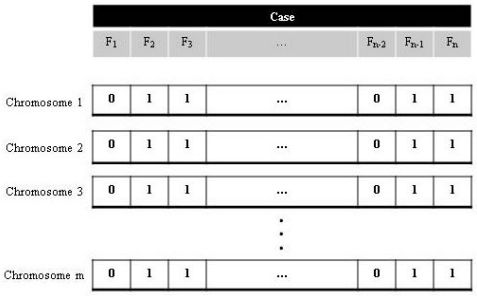


Fig.2 Chromosome Representation

### *C. Fitness function*

The goal of the proposed model is selecting the best subset of features that can produce the highest classification accuracy for diagnose the breast cancer. Therefore, the best subset of features should be selected. For selection the best subset, a function is needed to evaluate the result of each subset of features (Chromosome). In this model, CBR is employed for fitness function.

Each chromosome is sent to the CBR engine. Each gene on chromosome is a bit string value. It determines that the corresponding feature should be used in CBR process or not. For calculating fitness value of each chromosome based on CBR a test set is needed. Hence, the case base is divided into two sets. One of them is training set that is called case base and another is test set. Also, training set divided into training and validation sets. In chromosome evaluation step, all cases of validation set are given to CBR engine one by one. Each case is represented with n features. Each feature has a numerical value, so we can say that each case is a numerical vector. For every case from validation set, CBR searches the case base and retrieve the most similar case. For doing this step the Euclidian distance formula is used [27]:

$$d(a,b) = \sqrt{\sum_{i=1}^{k} (a_i - b_i)^2} \quad (1)$$

In which a and b are the validation case and a case from case base, k denotes the number of cases in case base, ai and bi are ith feature of a and b, d is the distance between two vector a and b. The larger the distance, the smaller the similarity is. Therefore, the case with minimum value of distance is retrieved. Afterward, the solution of retrieved case is selected as a solution for the case. Then, the given result compares with the real result. For all of the cases in validation case, these steps are repeated. Finally, the mean value of the fitness values is returned as evaluation result of a chromosome.

### *D. Genetic Operators*

For generate new population in order to maximize the fitness value, some genetic operators such as selection, mutation and crossover is used. After calculating fitness value for each chromosome by fitness function, there is a list of chromosomes with their fitness value. A selection operator selects top chromosomes based on their fitness value. Crossover exchanges substring from pairs of chromosome to generate two new chromosomes. In the proposed model two-point crossover is used. In mutation, selected genes are inverted. Mutation prevents the search process from falling into local maxima [22].By these operators, the new populations are generated and the fitness values for chromosomes in each population are calculated. This process continues until stopping criteria is satisfied. At the end, the chromosome with maximum fitness value is selected. The selected chromosome denotes which features are appropriate to classification process.

## II. Results

### *A. Dataset*

For evaluating the model, Wisconsin Diagnostic Breast Cancer (WDBC) Dataset is used. Each record of this dataset is represented with 30 numerical features. Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. The diagnosis of each record is “benign” or “malignant”. This dataset contains 569 instances. 357 instances are benign and 212 malignant. There is no missing value in the dataset.

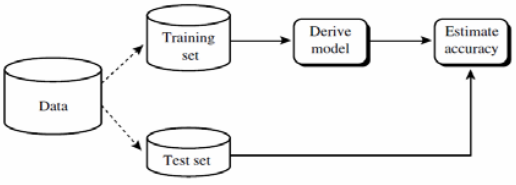


Fig.3 Holdout Method

### *B. Evaluation*

For estimating the accuracy rate of the proposed model, holdout method [4] is employed. In holdout method, dataset is divided into two sets. One of them is training set that is used for model training and another is test set that is used for estimating accuracy of the model. So, 80% of data is allocated to training set and the remaining 20% is allocated to test set. The process of holdout method is shown in figure 3.
The CBR classifier evaluated before feature selection and after it. The comparison results of the proposed model and the other models are shown in table 1.
The proposed Feature selection algorithm selected 12 features. Thus, CBR classifier has the best outcome, when these 12 features are used.

## III. Conclusion

In this paper, a case-based feature selection using GA for selecting the best subset of features for breast cancer diagnosis system proposed. GA used to search the problem space to find all of possible subsets of features and CBR employed to estimate fitness value of each chromosome. At the end, the best subset of features denoted.

For evaluating the proposed model, hold out method used. In order to holdout method, the dataset divided into two sets: training set and test set. The training set used for training model and test set used for estimating the accuracy of the model.

The goal of this paper was to find out that the proposed model is comparable with the other models or not, because this study is a pre study for selecting the optimal features from a real breast cancer database. As shown in table 1 the proposed model is comparable with the other models on Wisconsin breast cancer dataset. There is little difference between the proposed model and the others.

TABLE I
COMPARISON ACCURACY RATES BETWEEN PROPOSED MODEL AND OTHER MODELS

| Name | Algorithm | Accuracy with all features | Accuracy after feature selection |
|---|---|---|---|
| Bacauskiene [28] | Neural Network | 97.36 | 98.24 |
| Prasad [29] | ACO-SVM | 94.55 | 98.83 |
| Prasad [29] | GA-SVM | 94.55 | 98.95 |
| Prasad [29] | PSO-SVM | 94.55 | 100 |
| **Proposed Model** | **CBR-Genetic** | **94.74** | **97.37** |

## References

[1] I. Harirchi, et al., “Breast cancer in Iran: a review of 903 case records,” Public Health, 2000. 114(2): p. 143-145.
[2] T. Subashini, V. Ramalingam, and S. Palanivel, “Breast mass classification based on cytological patterns using RBFNN and SVM,” Expert Systems with Applications, 2009. 36(3): p. 5284-5290.
[3] R.A. Miller, “Medical diagnostic decision support systems - past, present, and future,” Journal of the American Medical Informatics Association, 1994. 1(1): p. 8.

[4] J. Han, and M. Kamber, “Data mining: concepts and techniques,” 2006: Morgan Kaufmann.
[5] R. Kohavi, and G.H. John, “Wrappers for feature subset selection,” Artificial intelligence, 1997. 97(1-2): p. 273-324.
[6] Y. Yuling, “A Feature Selection Method for Online Hybrid Data Based on Fuzzy-rough Techniques,” 2009: IEEE.
[7] N. Abe, et al., “A divergence criterion for classifier-independent feature selection,” Advances in Pattern Recognition, 2000: p. 668-676.
[8] M. Dash, and H. Liu, “Feature selection for classification,” Intelligent data analysis, 1997. 1(3): p. 131-156.
[9] R. Jensen, and Q. Shen, “Computational intelligence and feature selection: rough and fuzzy approaches,” IEEE Press Series On Computational Intelligence, 2008: p. 340.
[10] I. Guyon, and A. Elisseeff, “An introduction to variable and feature selection,” The Journal of Machine Learning Research, 2003. 3: p. 1157-1182.
[11] M. Sun, et al. “A GA-Based Feature Selection for High-Dimensional Data Clustering,” 2009: IEEE.
[12] C.H. Yang, et al., “A Novel GA-Taguchi-Based Feature Selection Method,” Intelligent Data Engineering and Automated Learning–IDEAL 2008, 2008: p. 112-119.
[13] I.S. Oh, J.S. Lee, and B.R. Moon, “Hybrid genetic algorithms for feature selection,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2004: p. 1424-1437.
[14] P. Zhang, B. Verma, and K. Kumar, “Neural vs. statistical classifier in conjunction with genetic algorithm based feature selection,” Pattern Recognition Letters, 2005. 26(7): p. 909-919.
[15] J.H. Hong, and S.B. Cho, “Efficient huge-scale feature selection with speciated genetic algorithm,” Pattern Recognition Letters, 2006. 27(2): p. 143-150.
[16] R. Leardi, and A. Lupiáñez González, “Genetic algorithms applied to feature selection in PLS regression: how and when to use them,” Chemometrics and Intelligent Laboratory Systems, 1998. 41(2): p. 195-207.
[17] M., Mitchell, “An introduction to genetic algorithms,” 1998: The MIT press.
[18] J.H. Holland, “Adaptation in natural and artificial systems,” 1992: MIT Press Cambridge, MA, USA.
[19] A. Aamodt, and E. Plaza, “Case-based reasoning: Foundational issues, methodological variations, and system approaches,” AI communications, 1994. 7(1): p. 39-59.
[20] H. Ahn, K. Kim, and I. Han, “A case-based reasoning system with the two-dimensional reduction technique for customer classification,” Expert Systems with Applications, 2007. 32(4): p. 1011-1019.
[21] H. Ahn, K. Kim, and I. Han, “Hybrid genetic algorithms and case based reasoning systems for customer classification,” Expert Systems, 2006. 23(3): p. 127-144.
[22] H. Ahn, and K. Kim, “Bankruptcy prediction modeling with hybrid case-based reasoning and genetic algorithms approach,” Applied Soft Computing, 2009. 9(2): p. 599-607.
[23] K.J. Kim, “Toward global optimization of case-based reasoning systems for financial forecasting,” Applied Intelligence, 2004. 21(3): p. 239-249.
[24] G.R. Beddoe, and S. Petrovic, “Selecting and weighting features using a genetic algorithm in a case-based reasoning approach to personnel rostering,” European Journal of Operational Research, 2006. 175(2): p. 649-671.
[25] J. Jarmulak, S. Craw, and R. Rowe, “Genetic algorithms to optimise CBR retrieval,” Advances in Case-Based Reasoning, 2000: p. 159-194.
[26] E. Golobardes, X. Llor , and M. Salamó, “Computer aided diagnosis with case-based reasoning and genetic algorithms,” Knowledge-Based Systems, 2002. 15(1-2): p. 45-52.
[27] Y. Avramenko, and A. Kraslawski, Case Based Design. Applications in Process Engineering, 2008: p. 51-108.
[28] M. Bacauskiene, and A. Verikas, “Selecting salient features for classification based on neural network committees,” Pattern Recognition Letters, 2004. 25(16): p. 1879-1891.
[29] Y. Prasad, K. Biswas, and C. Jain, “SVM Classifier Based Feature Selection Using GA, ACO and PSO for siRNA Design,” Advances in Swarm Intelligence, 2010: p. 307-314.