IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-4, April 2019 E-ISSN: 2347-2693

Breast Cancer Detection using Genetic Algorithm with Correlation based Feature Selection: Experiment on Different Datasets

Shivangi Singla[1*], Pinaki Ghosh[2], Uma Kumari[3]

[1,2]Dept. of Computer Science and Engineering, Mody University of Science and Technology, Lakshmangarh, Rajasthan, India

Corresponding Author: shivangisingla306@gmail.com

DOI: https://doi.org/10.26438/ijcse/v7i4.406410 | Available online at: www.ijcseonline.org

Accepted: 15/Apr/2019, Published: 30/Apr/2019

Abstract- Breast cancer is second leading invasive cancer causes death after lung cancer. The accurate diagnosis is a very crucial aspect of breast cancer treatment. For this purpose, data mining techniques guide doctors in correct decision-making for diagnosis. This paper demonstrates various data mining methods for breast cancer diagnosis. The proposed algorithm is distinguished into two sections. First section consists of feature selection methods to reduce the computational complexity, as genetic algorithm is used to eliminate the irrelevant features from the dataset and second section describes different classification algorithms named Multilayer Perceptron, Random Forest, and Naive Bayes classification to determine whether breast cancer is malignant or benign type. The proposed algorithm is applied to four datasets of Wisconsin Breast Cancer Dataset and at last comparison is made between various classification algorithms to achieve highest classification accuracy.

Keywords- Feature Selection, Genetic Algorithm, Multilayer Perceptron, Random Forest, Naive Bayes.

I. INTRODUCTION

Cancer arises when there is uncontrolled, unacceptable and uncoordinated growth of cells in the body which further replicate into additional cells and harmful for the body. Breast cancer occurs when that uncontrolled cell growth in breast portion [1]. Detection of breast cancer at the initial stage is crucial for health care. To reduce the expansion of cancer tissue of other parts of the body, early detection and treatment is very essential. Breast tumors are distinguished in two types named malignant tumor and benign tumor. Benign tumor is non-cancerous and non-dangerous tumor, which grows only in one part of the body and can be treated well with medicines or surgery. Malignant tumor is defined as the cancerous tumor as the cancerous cell grows in another part of the body. So determination of the type of breast tumor is a very important task and it should be reliable and accurate for the sake of life's quality. Various diagnostic methods are adopted such as mammography, biopsy, ultrasound, fine needle aspiration cytology and thermography [2]. These techniques are expensive and invasive and the accuracy rate is also not so very high.

Numerous methods and technologies are developed for analyzing and collecting the data, but it is very challenging task for physicians to understand each and every particularized cancer feature from very huge volume of data. Therefore, data mining techniques have become a valuable representative for physicians. Data mining is the process which extracts the valuable and informative data from large dataset [3]. With the help of data mining techniques, an intelligent system for detection is discovered. The detection system considers three phases as a first phase determines a distinct disease dataset, second phase describes the reduction of feature for high dimensional content and in third phase classification accuracy is developed to test that system is efficient or not.

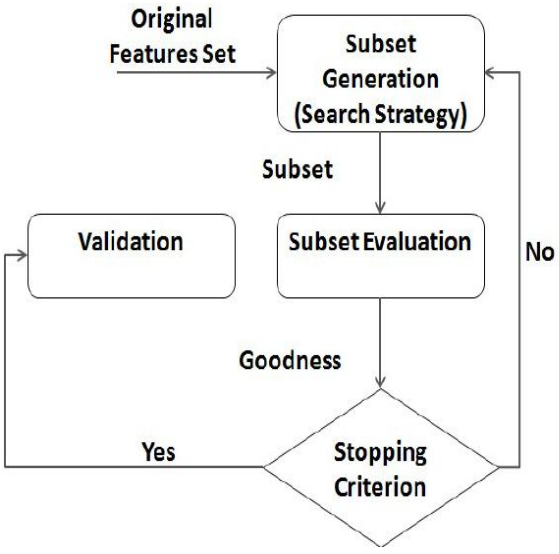

Fig.1- Steps of Feature Selection

With the advancement of various techniques, various datasets consist of redundant and many irrelevant attributes. These attributes deteriorate efficiency of the algorithm as irrelevant attributes degrade the classification accuracy. Relevant attributes contribute to classification accuracy only and useful in reducing the computational cost and time and helps in making better the classification model's performance. Dimension reduction can be classified by two techniques, namely, extraction of feature and selection of feature.

Feature extraction is the process extracting the subset of new features by combining the old features. It is conversion of high dimensional feature into lower dimensional features. Feature selection is the selection of relevant or useful features from the existing features. Feature subset is expected to be good so that it can gratify subsequent characteristics. First, irrelevant features generate the large search space. Second, subset of feature which is selected commits the complete information needed to discriminate patterns. Third, helps to reduce the size of dataset. Fourth, on the dimensionality of the subset, the computational time and cost depend [4]. Feature selection consists of four steps described in Fig.-1.

Rest of the paper is organized as follows, Section I describes the introduction about the breast cancer and feature extraction technique, Section II contains the related work done by the various researchers, Section III contains dataset description used in the experimentation work, Section IV includes the architecture of the proposed work with detailed description including flow chart, Section V contains the results with the complete discussion and Section VI concludes the research work with future scope.

II. RELATED WORK

Shokoufeh Aalaei [5] proposed an algorithm which improves the classification accuracy from the previous researches. The proposed algorithm used the genetic algorithm for the feature selection process for the diagnosis of breast cancer. Three types of datasets are used named as Wisconsin Breast Cancer dataset, Wisconsin Diagnosis Breast Cancer dataset, Wisconsin Prognosis Breast Cancer dataset. The number of features are selected from these datasets are respectively WBC-4, WDBC-14, WPBC-16. PS-classifier, ANN, GA-classifier used to examine the feature selection process. The results of proposed algorithm are as in WBC, PS-classifier gives the highest classification accuracy. In WPBC and WDBC, highest accuracy is achieved by ANN classification algorithm. Accuracy, sensitivity and specificity are calculated with and without feature selection method. This study shows that the classification accuracy can be improved by selecting subset of relevant features.

Kavitha C.R [6] presented two feature selection methods named as Information gain and forward selection (IGfwS) and Recursive feature elimination with SVM (SVMRFE). After this, a hybrid approach is made using rough set theory (RST) with both the feature selection methods and described as RST+IGfwS and RST+SVMRFE. It uses four datasets of different diseases named Dermatology, Hypothyroid, Liver Disorder and Wisconsin Breast Cancer. Later comparison is made for all the four methods. The main objective of this study is accomplished multiclass and binary classification more precisely and flawlessly with the help of a reduced set of features selected. From the study, it is examined that the best classification accuracy is given by SVMRFE and RST+SVMRFE methods and different classification algorithm are used such as random forest, IBK, J48 and Jrip. From these classification algorithms, random forest gives the best classification accuracy with all four datasets.

B. Tamilvanan [7] demonstrated two feature selection processes as genetic algorithm and random search and for evaluating the feature, correlation based feature selection is used. The dataset used is Wisconsin Breast Cancer dataset contains 10 attributes and 286 instances. Different classification algorithms are proposed to be a Naive Bayes classification, Multilayer Perceptron, Random Forest and Sequential Minimal Optimization for determining various classification factors like accuracy, precision, specificity, sensitivity and time. Firstly, it discovers which classification algorithm gives the highest accuracy without using the feature selection method. Further, it uses the genetic feature selection method on WBC dataset and applied classification algorithm and finds that Naive Bayes classification gives better results amongst results. Later, it applied same with the random search method and again Naive Bayes describes the better results. Finally, it comes to an end by examining that Genetic Algorithm based correlation feature selection and Naive Bayes algorithm presents the best results for breast cancer diagnosis.

III. DATASET DESCRIPTION

Four different datasets are proposed for the study. The datasets named Wisconsin Breast Cancer (WBC), Wisconsin Diagnostic Breast Cancer (WDBC), Wisconsin Prognostic Breast Cancer (WPBC) and inbuilt dataset in WEKA tool. The description of datasets is given in table 1. The inbuilt dataset is the dataset used by the tool WEKA and described in it with 10 attributes and 286 instances with output class of recurrent and non-recurrent type. The Wisconsin Breast Cancer dataset contains 699 instances and 11 attributes with benign and malignant type output class. In this dataset, the attributes are valued from 1-10 considering 1 being the least abnormal state and 10 being the highest abnormal state. The Wisconsin Diagnosis Breast Cancer dataset consist 32 attributes and 569 instances for diagnosis of breast cancer. In this dataset, various factors like area, perimeter, etc. are

considered and class is examined as either benign or malignant type. The Wisconsin Prognosis Breast Cancer dataset gives 198 instances and 35 attributes with recurrent and non-recurrent type of output class and there are 4 cases which are missing. The dataset WPBC and WDBC have almost same attributes, but WPBC have two more attributes named Lymph node and Tumor size. The attributes of WDBC and WPBC are extracted from digital images of breast mass using the technology named Fine Needle Aspirate (FNA).

Table 1: Dataset's Characteristics

Dataset Name	Inbuilt(WEKA)	WBC	WDBC	WPBC
No. of instances	286	699	569	198
Input Attributes	9	10	31	34
Output Classes	2	2	2	2
Total No. of Attributes	10	11	32	35

IV. PROPOSED WORK

In this section, discussion is made on our proposed work which includes feature selection using genetic algorithm and various classification algorithms. The proposed algorithm is described in Fig. 2 with detailed description.

Feature Selection

Feature selection is the process to remove irrelevant features from the large set of features or attributes. Feature selection algorithms are planned with three categories, namely, Filter model, Wrapper model and Hybrid model. Filter model designed to evaluate general properties of data and subsets of features are selected without including data mining methods. Wrapper model relies on one data mining algorithm for the feature selection to the improve the mining performance. Hybrid model combines both the techniques for better evaluation.

CFS Feature Evaluator

CFS is the type of filter model feature selection. It stands for Correlation Based Feature Selection. Highly correlated features or attributes are avoided. CFS deletes and adds one feature at a time. The irrelevant features are removed out by considering the Pearson's correlation.

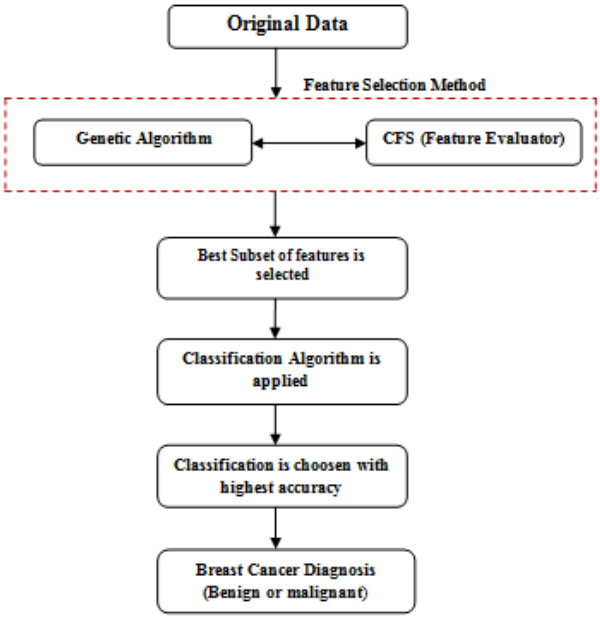

Fig. 2- Steps of proposed algorithm

Genetic Algorithm

Different search methods are used to find a good quality of subset. In our proposed work, Genetic algorithm is used as a search method to better subset of features and quality of extracted subset of features is examined by feature evaluator (CFS). Number of generations, probabailities of crossover and mutation and population size are the factors of genetic algorithm. Search point is stated by generating the list of attributes as initial population. In other words, CFS is used to find the finite number of subsets of features and from that subset, the features are selected which gives the better quality of subset for which genetic algorithm is used. As the genetic algorithm uses various parameters, the chromosomes are used as a mask of features and features which are selected are considered as 1 and which are not are considered as 0. It is considered that CFS is used to calculate the fitness function for genetic algorithm. It assits, ranking of the distinctive attributes using correlation coefficients. It is observed that the highly correlated attributes are ignored. For improving the quality of search methods, fitness function should be high. Higher the value of fitness function, lower will be the correlation between the attributes.

Classification Algorithms

Classification algorithms are applied to determine accuracy, time to execute the algorithm and performance of the prescribed work. Here in our experiment, various classification algorithms are used, namely Naive Bayes algorithm, Random Forest and Multilayer Perceptron. The best subset of features are selected using genetic algorithm and CFS evaluator. Now, accuracy is to be calculated using

classification algorithms and identifying the better algorithm for classification giving highest accuracy for the prescribed algorithm. A comparison is made on these algorithms with all four datasets described.

V. RESULTS AND DISCUSSION

The proposed approach for feature selection is applied using genetic algorithm as search method and CFS as feature evaluator. Table 2 describes the selected features for all four datasets. In inbuilt dataset of breast cancer from WEKA tool, 5 features or attributes are selected from 10 attributes which are considered to be the most relevant set of features. In WBC, 9 attributes are selected out of 11 attributes. In WDBC and WPBC, 14 and 3 attributes are selected respectively.

Table 2: Features selected for all four datasets

Dataset Name	Features selected using Genetic Algorithm
Inbuilt(WEKA)	3,4,5,6,9
WBC	2,3,4,5,6,7,8,9,10
WDBC	1,3,9,10,11,15,23,24,25,26,27,28,29,30
WPBC	2,20,26

After getting best subset of features, classification algorithms are applied to all the datasets to determine the classification accuracy and time taken by the algorithm to occur. The comparison is made between three algorithms, namely Naive Bayes, Random Forest and Multilayer Perceptron to evaluate better algorithm amongst these for the diagnosis of breast cancer either it is benign or malignant so that the treatment of the patient should be accurate and saves the quality of life.

Table-3: Determining the classification accuracy using various algorithms on different datasets

Classification Algorithms	Inbuilt dataset	WBC	WDBC	WPBC
Naive Bayes	74.8%	96.13%	94.20%	78.78%
Multilayer Perceptron	86.36%	99.28%	97.89%	81.81%
Random Forest	85.31%	99.4%	99.5%	99.6%

From table 3, it is found that Multilayer Perceptron algorithm gives better result for inbuilt dataset given in WEKA tool. For WBC, WDBC and WPBC, highest classification accuracy is achieved by the Random Forest algorithm.

Comparison of Our Work with Previous Work

Our proposed work is described as the classification accuracy is evaluated using genetic feature selection method and CFS as feature evaluator and by applying different classification algorithms as described in table 3. Now a comparison with previous work and our proposed work is to be made.

Table-4: Comparison of previous work with the proposed work for WDBC dataset

Feature Selection and Classification Algorithm	Classification Accuracy
RST+SVMRFE and Random Forest**[6]**	96%
SVMRFE and IBK algorithm**[6]**	95.8%
PS-Classifier and Genetic algorithm**[5]**	96.6%
GA With CFS evaluator and Random Forest **[proposed work]**	99.5%

Table-5: Comparison of previous work with the proposed work for inbuilt dataset

Feature Selection and Classification Algorithm	Classification Accuracy
Genetic Algorithm and Naive Bayes**[7]**	72%
Genetic Algorithm and MLP**[7]**	71%
Random Search and Naive Bayes**[7]**	77%
GA With CFS evaluator and MLP**[proposed work]**	86.36%

Table-6: Comparison of previous work with the proposed work for WPBC dataset

Feature Selection and Classification Algorithm	Classification Accuracy
Fisher Filtering and Naive Bayes **[9]**	75.25%
PS-Classifier and ANN **[5]**	79.2%
PS-Classifier and Genetic Algorithm **[5]**	78.2%
GA With CFS evaluator and Random Forest **[proposed work]**	99.6%

Table-7: Comparison of previous work with the proposed work for WBC dataset

Feature Selection and Classification Algorithm	Classification Accuracy
Genetic Algorithm and Logistic Regression **[8]**	98.45%
Genetic Algorithm and Decision Tree **[8]**	94.02%
PS-Classifier and	96.6%

ANN **[5]**	
GA With CFS evaluator and Random Forest **[proposed work]**	99.4%

VI. CONCLUSION

Feature selection is the process of determining the relevant features from huge sets of data to reduce the computational complexity and improve the accuracy or say the performance of the algorithm so that accurate and efficient results are to be evaluated. In our proposed work, genetic algorithm is selected as the search method for deciding the best subset of features and CFS is preferred as feature evaluator. More correlated attributes less will be the chances to be chosen and smaller the fitness functions. For electing the best subset of features, fitness function should be tabbed as high as possible. Later on, distinctive classification algorithms are selected for examining the classification accuracy. The proposed algorithm is applied on four datasets where one dataset is inbuilt given by WEKA tool and other are preferred from Wisconsin Breast Cancer datasets. Finally, it is concluded that Multilayer Perceptron algorithm is superior for inbuilt dataset and Random Forest algorithm excelling for another three Wisconsin Breast Cancer Datasets.

In future work, new feature selection methods with hybridizing different methods can be examined for better performance. Other diverse classification algorithms can be considered for better classification accuracy. Despite this, feature selection methods have disadvantages also. The feature selection method sometimes does not give better results for multiclass datasets and also takes more time and space which is major issue of concern. So, taking into knowledge, the best feature selection algorithm is to be considered which works better for multiclass datasets and binary datasets also.

REFERENCES

[1] G. Devi , et at. *"Breast Cancer Prediction System using Feature Selection and Data Mining Methods"*, International Journal of Advanced research in Computer Science, Vol. **2**, Issue **1**, pp. **81-87**, **2011**.

[2] R. Nithiya, et al. *"A data Mining Techniques for Diagnosis of Breast Cancer Disease"*, World Applied Sciences Journal, 29 (Data Mining and Soft Computing Techniques), pp. **18-23**, **2014**.

[3] B. Zheng, et al. *"Breast Cancer Diagnosis based on Feature Extraction using a Hybrid of K-means and Support Vector Machine Algorithms"*. Expert Systems with Applications, Elsevier, **2013**.

[4] C. Lu, et al. *"An Intelligent System for Lung Cancer Diagnosis Using a New Genetic Algorithm Based Feature Selection Method*". Journal of Medical Systems, Springer Science, Vol. **38**, pp. **88-97**, **2014**.

[5] S. Aalaei, et al. "*Feature Selection using Genetic Algorithm for Breast Cancer Diagnosis: Experiment on three Different Datasets*". Iran Journal of Basic Medical Sciences, Vol. **19**, Issue **5**, pp. **476-482**, May **2016**.

[6] C.R., K., T, M., *"Feature Selection Methods for Classification: A Comparison*". International Journal of Research in Engineering and Technology, Vol. **06**, Issue **06**, pp. **130-137,** June **2017**.

[7] B. Tamilvanan, V. M. Bhaskaran, "*New Feature Selection Techniques Using Genetics Search and Random Search Approaches for Breast Cancer*". Biosciences Biotechnology Research Asia, Vol. **14**, Issue **1**, pp. **409-414**, **2017**.

[8] E. Alickovic, et al. "*Breast Cancer Diagnosis using Genetic Algorithm Feature Selection and Rotation Forest*". The Neural Computing and Applications, Springer, **2015**.

[9] D. Dumitru, et al. "*Prediction of Recurrent Events In Breast Cancer using the Naive Bayesian Classification*". Mathematics and Computer Science, Vol. **36**, Issue **2**, pp. **92-96**, **2009**.

[10] T. Karthikeyan, et al. "*Genetic Algorithm based CFS and Naive Bayes Algorithm to Enhance the Prediction Accuracy*". Indian Journal of Science and Technology, Vol. **8**, Issue **27**, **2015**.

[11] S. Vanaja, et al. "*Analysis of Feature Selection Algorithms on Classification: A Survey*". International Journal of Computer Applications, Vol. **96**, Issue **17**, pp. **1-8**. June **2014**.

[12] R. Tiwari, et al. *"Correlation Based Attribute Selection Using Genetic Algorithm*". International Journal of Computer Applications, Vol. **4**, Issue **8**, pp. **28-34**, August **2010**.

[13] R. Suji, S. P. Rajagopalan."*Multi-Ranked Feature Selection Algorithm for Effective Breast Cancer Detection*". Biomedical Research, pp. **S99-S102**, **2016.**

[14] D. Lavanya et al. "*Analysis of Feature Selection with Classification: Breast Cancer Datasets*". Indian Journal of Computer Science and Engg. Vol. **2**, Issue **5**, pp. **576-563**, Oct-Nov **2011**.

[15] Z. Karanpunar, et al. "*Breast Cancer Diagnosis via Data Mining: Performance Analysis of Seven Different Algorithms*". Computer Science and Engineering: An International Journal, Vol. **4**, Issue **1**, pp. **35-46**, Feb. **2014**.

